Water Integrity Risk Index

Online Tool Manual

Table of contents

What is WIRI?

The Water Integrity Risk Index (WIRI) is a tool for measuring integrity in the water and sanitation (W&S) sector at the city level (Fazekas et al. 2020). It relies on public procurement and survey data in order to calculate a composite indicator which ranges between 0 and 100, where **cities with scores closer to 100 have lower risks of corruption**.

Unlike existing indexes, WIRI incorporates both Big Data (public procurement) and traditional statistical methods (surveys) to develop a multidimensional indicator based on **objective data** rather than *perceptions* of corruption. In order to get as clear a picture as possible of corruption and integrity trends, WIRI is calculated at the level of **cities**. This enables the assessment of smaller changes in integrity both across specific localities and over time.

The data that is uploaded into this online tool is classified into three pillars: a) *operations* for day-to-day W&S related activities, b) *investments* for large scale projects, and c) *interactions* for direct engagement with water utilities.

The WIRI tool requires at least one data set on **public procurement** in the water and sanitation sector (i.e. government contracts for W&S related goods and services). In addition, users have the **option** to upload a survey data set on direct experiences with corruption and a custom keywords data set which classifies contracts into one of the three pillars of WIRI. There are examples of all these data sets available in the WIRI tool website.

This manual will guide users through the process of constructing data sets in the **specific format** that is required to upload them to the control panel of the WIRI Tool. Once they have been created and uploaded correctly, the tool will automatically calculate the WIRI score for each city and display a summary of the results in a dashboard.

What is Public Procurement?

The main data source for WIRI is public procurement contracts. Public procurement refers to the process by which government departments or agencies purchase goods or services from the private sector. This system ensures that public sector needs are met in a transparent, efficient, and fair manner, often contributing to economic growth and fostering public trust. The main stages of procurement are:

1. **Planning**: Identification of needs and budget allocation.
2. **Tender Announcement**: Publishing the need for a specific good, service, or work.
3. **Bidding**: Private sector companies submit their proposals or bids.
4. **Evaluation**: Bids are assessed based on predetermined criteria.
5. **Award**: The contract is awarded to the winning bidder.
6. **Contract Implementation**: Delivery of goods or services.
7. **Review and Audit**: Ensuring compliance with the contract and assessing performance.

Public procurement data sets are structured collections of data that capture the details of contracts awarded by government agencies or departments to private entities. These data sets are designed to ensure transparency, accountability, and efficiency in the use of public funds. Similarly, procurement data is often used to measure corruption risks in the W&S sector and beyond (Adam et al. 2020, 2023).

The main variables often include the **winner's name**, denoting the contractor awarded the bid, and **bids count**, indicating the competitiveness of the bidding process. The timeline of the procurement process is captured through the **first call date**, **bid deadline**, and **award date**. The **procedure type** reveals the methodology behind the contract award, while the **contract title** provides a snapshot of the contract's essence. Crucial details about the purchasing agency are captured in the **buyer's name** and **buyer city**. Lastly, the **final value** provides a quantitative measure of the contract's monetary worth. These variables are common across different procurement data sets, though some may not report them all and many data sets have dozens of additional variables.

WIRI uses these public procurement variables to identify red flags of corruption and calculates a Corruption Risk Indicator (CRI) (Fazekas, Tóth, and King 2016) index in the background. Table 1 defines how these red flags of corruption are defined and how they are used to construct the CRI, which in turn is integrated into the WIRI composite indicator. The classification of each contract into a risk category is done automatically by the WIRI tool based on procurement data imputed by users. The following section specifies how procurement data sets must be structured so that the WIRI tool can perform these background calculations and output the WIRI composite index.

Table 1: Public Procurement Corruption Risk Indicators

Indicator	Definition
Length of decision period	100=LENGTH OF DECISION PERIOD IS UNRELATED TO CORRUPTION RISKS, 0=LENGTH OF DECISION PERIOD OR MISSING DECISION PERIOD IS RELATED TO CORRUPTION RISKS
Call for tenders publication	100=CALL FOR TENDER PUBLISHED IN OFFICIAL JOURNAL, 0=OTHERWISE
Length of advertisement period	100=LENGTH OF ADVERTISEMENT PERIOD IS UNRELATED TO CORRUPTION RISKS, 0=LENGTH OF ADVERTISEMENT PERIOD OR MISSING DECISION PERIOD IS RELATED TO CORRUPTION RISKS
Procedure type	100=OPEN, 0=NON-OPEN
Single bidder contract	100=MORE THAN 1 BID RECEIVED, 0=1 BID RECEIVED

Public Procurement Data Set

This section specifies the format in which the **required** public procurement data set must be created in order to be used in the WIRI tool. The procurement data set should be organized in a CSV format (not as an Excel file) with the variables outlined below. Each row should represent one contract and each variable should be stored in a column. Table 2 shows an example of a correctly formatted procurement data set. The names of the variables and the format of the cells should be exactly as specified.

1. contract_title: The detailed description or title of the contract. This column should consist of text. For example: "Construction materials for drainage systems."
2. buyer_name: The name of the entity or agency that is buying or commissioning the service or product. For example: "Water Utility Services Company"
3. winner_name: The name of the entity or company that won the contract bid. For example: "Global Construction Materials Inc."
4. buyer_city: The city where the buying entity is located. For example: "Sheffield"
5. final_value: The final or settled monetary value of the contract. This should be a numerical value. Though it does not matter which specific currency the value is in, all values in the data set should be of the same currency. For example: 10555
6. bids_count: The total number of bids received for the contract. This should be an integer. For example: 8
7. bid_deadline: The date by which all bids should have been submitted in the following format: MM/DD/YYYY. For example: 06/25/2018
8. firstcall_date: The date the contract was first announced or made open for bidding in the following format: MM/DD/YYYY. For example: 06/13/2018
9. procedure_type: The type of procurement procedure used classified into one of the following three options: OPEN, RESTRICTED, NEGOTIATED.
10. award_date: The date the contract was awarded to the winning entity in the following format: MM/DD/YYYY. For example: 07/2/2018

contract_title	buyer_name	winner_name	buyer_city	final_value	bids_count	bid_deadline	firstcall_date	procedure_type	award_date
Water Main Upgrade	City of Chicago	Great Lakes Water Services	Chicago	4324	2	6/25/18	6/13/18	NEGOTIATED	7/2/18
Sewage Treatment Works	Detroit Transit Authority	Midwest Water Solutions	Detroit	324321	4	3/28/18	3/16/18	NEGOTIATED	4/9/18
Hydration Station Installation	City of Chicago	Great Lakes Water Services	Chicago	324454	3	9/5/18	8/22/18	NEGOTIATED	9/18/18
Wastewater Recycling	Department of Water Service	Midwest Water Solutions	Detroit	5425	5	12/11/18	10/31/18	OPEN	12/27/18
Drainage System Overhaul	Chicago Public Schools	AquaTech Utilities	Chicago	46532	6	11/22/18	11/9/18	NEGOTIATED	11/27/18
Water Purification Project	City of Detriot	AquaTech Utilities	Detroit	5436	1	10/9/18	9/26/18	NEGOTIATED	10/15/18
Sanitation Facilities Renewal	City of Chicago	Great Lakes Water Services	Chicago	24589	1	9/20/18	8/20/18	NEGOTIATED	9/25/18
Stormwater Management	Detroit Transit Authority	AquaTech Utilities	Detroit	564784	5	10/1/18	8/23/18	OPEN	10/10/18
Reservoir Maintenance	Department of Water Service	AquaTech Utilities	Chicago	224678	2	7/26/18	6/21/18	OPEN	8/1/18
Pipeline Rehabilitation	Cook County Hospitals System	Midwest Water Solutions	Chicago	232554	1	6/11/18	5/30/18	NEGOTIATED	6/13/18
Leak Detection Services	Cook County Hospitals System	Great Lakes Water Services	Chicago	5679	1	3/14/18	2/28/18	NEGOTIATED	3/16/18
Water Tower Refurbishment	Toledo Park Districts	Midwest Water Solutions	Toledo	2000	1	2/22/18	2/12/18	NEGOTIATED	2/23/18
Sewer Line Replacement	City of Chicago	AquaTech Utilities	Chicago	12498	2	5/16/18	3/22/18	OPEN	7/24/18
Flood Control System	Toledo Housing Authority	Great Lakes Water Services	Toledo	43457	3	10/5/18	8/10/18	OPEN	10/10/18
Water Meter Installation	Toledo Housing Authority	Midwest Water Solutions	Toledo	3546	1	6/11/18	8/10/18	OPEN	10/10/18

Table 2: Example of Public Procurement Data Set

:::

Public Procurement Data Sources

This section outlines some examples of online resources where you may find public procurement data sets. Once collected, the data sets should undergo a cleaning and structuring process in order to be used in the WIRI tool (see the example at the end of the manual). This data is typically publicly available and can be found across several sources.

1. Government Portals

- **U.S.**: Federal Procurement Data System (FPDS)
- **India**: Government e-Marketplace (GeM)
- **Peru**: Open Data Portal (CONOSCE)
- **Bangladesh**: e-Government procurement (e-GP)

2. International Organizations and Development Banks

- **World Bank**: The World Bank's Procurement Database
- **European Union**: Tenders Electronic Daily (TED)
- **Asian Development Bank**: Operational Procurement Database

3. Academic, NGOs and Third-party Platforms

- GovSpend
- Government Transparency Institute
- Harvard Dataverse
- Open Contracting Partnership. (OCP)

Building a Procurement Data Set for WIRI

1. Selecting a Data Cleaning Software

There are two main ways of constructing a WIRI formatted procurement data set from publicly available information: a) manually with the use of spreadsheet programs like MS Excel, and programatically with software like R or Python. The first option will require users to identify, rename and select relevant variables manually; whereas with the latter this can be done with a data cleaning pipeline (i.e. code). An example of a data-cleaning pipeline in R and Python can be found in the appendix of this manual.

2. Reading Original Data

Publicly available data sets on procurement (see section on procurement data) typically allow users to download Excel files (.xlsx) or CSV files (.csv) with contract-level data. These files can be opened with either spreadsheet software like Excel or tools like R and Python. Other procurement data sources, however, might require manual collection from online dashboards.

After an unprocessed procurement data set is loaded or collected, the key variables (e.g. `contract_title`) must be identified and, when relevant, renamed. This step ensures that the variables have the specific names required by the WIRI Tool (see section on variables). From the vast array of possible variables in the original data set, only 10 essential ones should be kept for WIRI.

3. Data Transformation

Once all variables have been selected and named correctly, they should be placed in the format required by the tool (e.g. dates as `MM/DD/YYYY`). Specific columns, such as the one indicating how a contract was awarded, should be reclassified into the categories required by the tool (e.g., `OPEN`, `RESTRICTED`, `NEGOTIATED`). For Excel users, this can be done using the `Format Cells` option or the `Replace` function for a highlighted column.

4. Focus on Water and Sanitation

The data should be filtered to retain only contracts related to water and sanitation. All contract titles, winner names, and buyer names that contain general W&S related keywords (e.g. water, sanitation, sewer, pipeline) should be kept. All other contracts should be filtered out. For Excel users, this can be done by highlighting the variables using the `Filter` function in the tab `Data` and selecting the option `Text Filters` from the dropdown menu. Some **examples of general W&S keywords** include, but are not limited to:

- water
- sanitation
- sewers
- drainage
- irrigation
- pipelines
- pumps
- hydration
- flood
- drinking

5. Data Quality Assurance

Consider removing observations with missing or incomplete information, especially regarding the location of the buyer. It is good practice to do some preliminary data analysis to determine the number of total observations per city in the data set.

6. Storing the Processed Procurement Data

Once all modifications are done, this cleaned and focused data set is saved as a separate CSV file. This file now contains a precise view of water and sanitation-related procurement contract which can then be uploaded into the WIRI Tool. It should look like the example in table 2.

Survey Data Set

This section specifies the format in which the **optional** survey data must be uploaded to the WIRI tool. In addition to public procurement data, the WIRI tool also offers the option of using survey data on direct experiences with corruption (e.g. anonymous admissions of bribery) in the W&S sector. There are several sources of this type of data collected from existing surveys, though users also have the option of using data collected from their own surveys.

Regardless of their source, in order to be used in the WIRI tool, survey data sets must have some specific characteristics:

- Responses should be recorded at the level of cities. For example, a survey on corruption in the W&S sector in the United States must have responses recorded for Chicago, and not for Illinois or the USA overall.
- When possible, response can also be recorded at the City, Year level (time-series-cross-section). For example, Chicago in 2007, 2008, and 2009.
- Each row in a survey must correspond to a city, or city year, and not to an individual respondent. This is meant to protect the anonymity of survey respondents.
- The survey data set should have exactly four variables, `buyer_city`, `year`, `n`, `bribes`. The variable for year must not be left empty even if there is only one round.
- The values for `buyer_city` and. `year` must overlap with values in those same variables in the procurement data set. For example, if we have survey data for Chicago in 2008, we must also have procurement data for Chicago in that same year. Similarly, locations must be named the same in both data sets. For example, if one is `Chicago` and the other is `City of Chicago`, then the data sets will not merge properly.
- The values for `n` should correspond to the total number of respondents from a given city in a given year, and `bribes` to the number of those that had a direct experience with corruption in the W&S sector. For example, if the survey included 288 respondents from Chicago in 2008, and 12 reported a direct experience with corruption in the sector, then the survey data set uploaded to the WIRI tool should look like the example below:

Table 3: Example of Survey Data

buyer_city	year	n	bribes
Chicago	2008	288	12
Toledo	2008	144	9
Detroit	2008	208	27

Examples of Third-Party Survey Data

- AfroBarometer: a pan-African, non-partisan research network that conducts public attitude surveys on democracy, governance, economic conditions, and related issues in more than 30 countries in Africa.
- Global Corruption Barometer: a worldwide public opinion survey conducted by Transparency International, which assesses general public views and experiences of corruption in various countries around the globe.

Custom Survey Considerations

Collecting surveys is a complex endeavor. Below are some general guidelines and considerations.

- Question Formulation: Frame the question to elicit clear, honest responses. Example: "Have you ever given or received any form of bribe or gift to obtain or provide water and sanitation services?"

- Response Options: Include straightforward options like 'Yes', `No`, and `Prefer not to say`. This allows for clear data analysis while respecting respondent privacy. Keep in mind that WIRI tool calculations will consider any response that is not explicitly `Yes` as a `No`.

- Define an adequate sample size: Identify the city's demographic segments - including age, income, and residential areas - to ensure a representative sample. Similarly, determine an adequate sample size to ensure statistical significance (Desu and Raghavarao 1990). Online tools such as this one can help you determine the adequacy of your sample size.

- Ethical Considerations and Informed Consent: Clearly explain the survey's purpose, confidentiality measures, and the respondents' right to withdraw at any time without consequences.Assure respondents that their identities will remain anonymous and their responses confidential. The WIRI tool will only input aggregate data (at the level of cities), but users must have their own data management protocols for respondent-level survey data they may collect.

- Data Collection and Analysis Data Integrity: Establish protocols to ensure data accuracy and prevent manipulation.

Survey data on direct experiences with corruption is sometimes collected within the scope of more general surveys by larger projects. Similarly, due to the sensitive nature of the topic, special care must be given to ensure that respondents feel comfortable communicating these experiences. Users that collect their own survey data on corruption in the W&S sector may consider the guidelines for corruption surveys outlined by the UN Office on Drugs and Crime (UNODC-INEGI 2018) or in academic literature (Reinikka and Svensson 2003). For a deeper overview on the topic, consider the guide by Transparency International. In addition, please consider sharing your aggregate survey data with other WIRI tool users by adding your survey data set to the WIRI Tool Data Repository.

Custom Keywords Data Set

This section specifies the format in which the **optional** custom keywords data must be uploaded to the WIRI tool. The WIRI composite index has three sub-indicators which are calculated by analyzing corruption risks in water-related public procurement contracts which are classified into three pillars: a) operations, b) investments, and c) interactions. These keywords identify if a contract can be categorized as related to ongoing **operations** (e.g. maintenance), **investments** in W&S infrastructure (e.g. drainage systems), or when a water utility (or similar institution) is acting as directly as the winner *or* buyer (e.g. the Chicago Water Company wins a contract by the Parks Authority to improve irrigation). The last point is take as a proxy for client-utility **interactions**. You may read more about the theory behind these categories in the WIRI working paper.

One of the main advantages of WIRI is its built-in flexibility. The *tool includes pre-defined general keywords for each of these pillars* in English and Spanish. However, since keywords can be very context-specific, the tool allows users to create their own list of keywords to classify contracts into one of the three pillars. For example, users that have a list of specific names of water utilities and service providers in the cities in their data sets will obtain better classifications for the interactions pillar. The more exhaustive the keywords, the more accurate the classifications. In addition, users that input data sets where data on contract titles or winner/buyer names are in more than one language (e.g. English and German) will obtain better results by defining their own list of keywords in both languages.

Defining Custom Keywords

Users have the option of uploading CSV files with custom keywords that will flag observations (contracts) in the public procurement data set into one of the three conceptual pillars. These keywords can be in any language, insofar as that language corresponds to the language of the public procurement data set. The keywords data set should have exactly three columns: `keywords_cui`, `keywords_op`, and `keywords_inv`. These correspond to keywords for Interactions, Operations, and Investments respectively. Below are some general considerations:

- Each row in the data set should correspond to a specific term.
- A single contract can be flagged into more than one pillar.
- Keywords can be specific or general.
- The tool will turn all imputed text into lower case letters.
- The choice of keywords will affect the number of observations in the WIRI score, so they should be considered carefully.

Table 4: Custom Keyword Example

keywords_cui	keywords_op	keywords_inv
Chicago Water Company	maintenance	pipeline
utility	service	sewage system
	clean	construction material

Based on the example in the table above, a contract with the title “`Construction material for a water treatment plant`” would be flagged as investments only, whereas “`maintenance services for water pipelines`” would be classified as both investments and operations. The code below shows and example of the string detection methodology used by the tool. For example, if a contract has the title `installation of water pipelines`, then the keyword `pipeline` will be detected and classify that contract as `investments` based on the definitions of table 4.

```
# String Detection Command (Text, Keyword)
str_detect("installation of water pipelines", "pipeline")
```

```
[1] TRUE
```

Troubleshooting

In order for the WIRI tool to work properly, CSV data sets must be formatted with specific variable names. The diagnostics box under the file upload will tell you if the files uploaded are correct or not. It may specify that the 'file is not a CSV' or that the 'variable names are incorrect'. To fix these issues, please consider the following:

- Make sure that you are uploading **CSV files** and **not Excel files**. CSV (.csv) files are very similar to Excel files (.xlsx), however, they are more portable and can only have a *single sheet*.
- Make sure that you are uploading the data sets to their specified box (e.g. procurement in the procurement upload section). If you are uploading survey data, then make sure that it is uploaded to the correct box. Otherwise, the app will tell you that: `Variable names are incorrect`.
- Verify that the names of the files are written exactly as specified. Otherwise, you will get the same error message: `Variable names are incorrect`.
- Verify that no variables are missing. If there is no information available for a variable, then the column should be filled with the text `NA` to indicate missing values.
- If you are uploading your own keywords, verify that all of the keywords in a custom keywords data set are specified in one of the three columns (one per pillar) of the data set.
- Make sure that columns with numeric values (e.g. `final_value`) contain numbers only. For example: `3144` is correct, whereas `3,144` is incorrect.
- The message `disconnected from server` will display after a few minutes of inactivity. If this occurs, users will have to re-upload their data sets and start the process once again.

Variable names should read **exactly** as follows, the order does not matter:

- For **Surveys**: `buyer_city`, `year`, `n`, `bribes`
- For **Procurement**: `contract_title`, `buyer_name`, `winner_name`, `buyer_city`, `final_value`, `bids_count`, `bid_deadline`, `firstcall_date`, `procedure_type`, `award_date`
- For **Custom Keywords**: `keywords_cui`, `keywords_op`, `keywords_inv`

Appendix: Building Procurement Data Sets

The following example builds a procurement data set based with information from the FCDO-funded (formerly DFID) project "Curbing Corruption in Government Contracting", which releases procurement data sets collected from national public procurement portals in 10 low and middle income countries: Chile, Colombia, India, Indonesia, Jamaica, Kenya, Mexico, Paraguay, Uganda and Uruguay.Specifically, we will use the data set on procurement in Mexico, which can be downloaded here. This section contains code that follows the steps outlined in the section "Building a Procurement Data Set".

Overall, this pipeline reads in a data set related to some form of procurement in Mexico, performs several transformations to clean and subset the data, and then saves the processed data to a new CSV file. Additionally, an empty survey data frame is created and saved to a separate CSV file.

Example in R (tidyverse)

```r
# Load necessary packages for data manipulation and date handling.
library(tidyverse)
library(lubridate)

# Read the CSV file into a data frame called 'mexico_dfid'.
mexico_dfid <- read_csv("dfid2_mx_210715_csv.csv")

# Start a data transformation pipeline on 'mexico_dfid'.
df_procurement <- mexico_dfid %>%
  # Rename the original columns for the required names.
  # This requires some prior data exploration.
  rename(
         winner_name = bidder_name,
         bids_count = tender_recordedbidscount,
         firstcall_date = tender_publications_firstcallfor,
         bid_deadline = tender_biddeadline,
         award_date = tender_contractsignaturedate,
         procedure_type = tender_proceduretype,
         contract_title = tender_title,
         buyer_name = buyer_name,
         buyer_city = buyer_city,
         final_value = bid_price
         ) %>%
  # Convert certain date columns from string format to Date format.
  # Recode the values in 'procedure_type' to new categories.
  mutate(
    firstcall_date = dmy(firstcall_date),
    bid_deadline = dmy(bid_deadline),
    award_date = dmy(award_date),
    procedure_type = case_when(
      procedure_type == "open auction" ~ "OPEN",
      procedure_type == "invitation (3 entities)" ~ "NEGOTIATED",
      procedure_type == "direct contracting" ~ "RESTRICTED",
      TRUE ~ NA_character_
    )
  ) %>%
  # Filter rows where either 'winner_name' or 'contract_title' contains specific  water keywords.
  filter(
    str_detect(winner_name, "agua+|acua+|hidr+") |
      str_detect(contract_title, "agua+|acua+|hidr+")
  ) %>%
  # Retain only the specific columns listed below.
  select(winner_name,
         bids_count,
         firstcall_date,
         bid_deadline,
         award_date,
         procedure_type,
         contract_title,
         buyer_name,
         buyer_city,
         final_value) %>%
  # Remove rows where 'buyer_city' is missing.
  filter(!is.na(buyer_city))

# Write the processed data frame 'df_procurement' to a new CSV file.
write_csv(df_procurement, "df_procurement.csv")

# Creating a dummy (empty) survey data set -------------------------------------

# Create an empty data frame called 'df_survey' with specific columns.
df_survey <- data.frame(
  buyer_city = NA,
  year = NA,
  n = NA,
  bribes = NA
)

# Write the empty 'df_survey' data frame to a new CSV file.
write_csv(df_survey, "df_survey.csv")
```

Example in Python

```
# Import necessary libraries for data manipulation and date handling.
import pandas as pd
from dateutil.parser import parse

# Read the CSV file into a DataFrame called 'mexico_dfid'.
mexico_dfid = pd.read_csv("dfid2_mx_210715_csv.csv")

# Rename the original columns for the required names.
# This requires some prior data exploration.
mexico_dfid.rename(columns={
    'bidder_name': 'winner_name',
    'tender_recordedbidscount': 'bids_count',
    'tender_publications_firstcallfor': 'firstcall_date',
    'tender_biddeadline': 'bid_deadline',
    'tender_contractsignaturedate': 'award_date',
    'tender_proceduretype': 'procedure_type',
    'tender_title': 'contract_title',
    'bid_price': 'final_value'
}, inplace=True)

# Convert certain date columns from string format to datetime format.
mexico_dfid['firstcall_date'] = mexico_dfid['firstcall_date'].apply(lambda x: parse(x).date())
mexico_dfid['bid_deadline'] = mexico_dfid['bid_deadline'].apply(lambda x: parse(x).date())
mexico_dfid['award_date'] = mexico_dfid['award_date'].apply(lambda x: parse(x).date())

# Recode the values in 'procedure_type' to new categories.
mexico_dfid['procedure_type'] = mexico_dfid['procedure_type'].replace({
    'open auction': 'OPEN',
    'invitation (3 entities)': 'NEGOTIATED',
    'direct contracting': 'RESTRICTED'
})

# Filter rows where either 'winner_name' or 'contract_title' contains specific water keywords.
mexico_dfid = mexico_dfid[mexico_dfid['winner_name'].str.contains('agua|acua|hidr', case=False, na=False) |
                          mexico_dfid['contract_title'].str.contains('agua|acua|hidr', case=False, na=False)]

# Retain only specific columns.
cols_to_keep = ['winner_name', 'bids_count', 'firstcall_date', 'bid_deadline', 'award_date',
                'procedure_type', 'contract_title', 'buyer_name', 'buyer_city', 'final_value']
mexico_dfid = mexico_dfid[cols_to_keep]

# Remove rows where 'buyer_city' is missing.
mexico_dfid.dropna(subset=['buyer_city'], inplace=True)

# Write the processed DataFrame 'mexico_dfid' to a new CSV file.
mexico_dfid.to_csv("df_procurement.csv", index=False)

# Creating a dummy (empty) survey data set -------------------------------------

# Create an empty DataFrame called 'df_survey' with specific columns.
df_survey = pd.DataFrame({
    'buyer_city': [None],
    'year': [None],
    'n': [None],
    'bribes': [None]
})

# Write the empty 'df_survey' DataFrame to a new CSV file.
df_survey.to_csv("df_survey.csv", index=False)
```

References

Adam, Isabelle, Mihály Fazekas, Alfredo Hernandez Sanchez, Peter Horn, and Nóra Regös. 2023. "Integrity Dividends: Procurement in the Water and Sanitation Sector in Latin America and the Caribbean." https://doi.org/10.18235/0004688.

Adam, Isabelle, Mihály Fazekas, Nóra Regös, and Bence Tóth. 2020. "Beyond Leakages: Quantifying the Effects of Corruption on the Water and Sanitation Sector in Latin America and the Caribbean." https://doi.org/10.18235/0002856.

Desu, M. M., and D. Raghavarao. 1990. "Power and Sample Size." In, 55–61. Elsevier. https://doi.org/10.1016/b978-0-12-212165-4.50009-7.

Fazekas, Mihály, István János Tóth, and Lawrence Peter King. 2016. "An Objective Corruption Risk Index Using Public Procurement Data." *European Journal on Criminal Policy and Research* 22 (3): 369–97. https://doi.org/10.1007/s10610-016-9308-z.

Fazekas, Mihály, Umrbek Allakulov, Alfredo Hernandez Sanchez, and Joshua Aje. 2020. "Water and Sanitation Sector Integrity Risk Index." *Unpublished*. https://doi.org/10.13140/RG.2.2.34948.96643.

Reinikka, Ritva, and Jakob Svensson. 2003. *Survey Techniques to Measure and Explain Corruption*. The World Bank. https://doi.org/10.1596/1813-9450-3071.

UNODC-INEGI. 2018. *Manual on Corruption Surveys*. Vienna: UNODC-INEGI Center of Excellence in Statistical Information on Government, Crime, Victimization; Justice.