Timeline

Description automatically generated

***Mohamed Ali Badran***

***May 2024***

***November 2024***

***KR***

Global Public Procurement Dataset Analysis

**MASTER THESIS NO. 2024: \_\_\_\_**

**College of Business and Economics**

**Department of Statistics &Business Analytics**

United Arab Emirates University

College of Business and Economics

Statistics &Business Analytics

Global Public Procurement Dataset Analysis

KR

2010048

This Capstone Project is submitted in partial fulfilment of the requirements for the degree in Master of Business (Business Analytics).

November 2024

**United Arab Emirates University Master Thesis**

**2024: XX**

© 2024 KR, UAE All Rights Reserved

Print: University Print Service, UAEU 2024

# Declaration of Original Work

I, KRR, the undersigned, a researcher and author of this report entitled “Analysis of the Global Public Procurement Dataset (GPPD)”, hereby solemnly declare that this is the original research work conducted by me. This report was prepared following thorough analysis of the GPPD dataset under the guidance and methodologies outlined in this document.

The findings, insights, and recommendations presented herein are the result of rigorous analysis of over 72 million contracts from 42 countries, covering the period from 2006 to 2021. This report has not previously formed the basis for the award of any academic degree, diploma, or similar title at any institution. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in this report have been properly cited and acknowledged in accordance with appropriate academic conventions.

I further declare that there are no potential conflicts of interest with respect to the research, data analysis, authorship, presentation, and publication of this report.

Student’s Signature:

Date:

# Advisory Committee

1. Advisor: Dr. Gurdal Ertek

Title: Associate Professor

Department of Statistics &Business Analytics (CBE)

College of Business and Economics

1. Member:

Title:

Department of

College of

1. Member:

Title:

Department of

College of

# Approval of the Master Thesis

This Master Thesis is approved by the following Examining Committee Members:

1. Advisor: Dr. Gurdal Ertek

Title: Associate Professor

Department of Statistics &Business Analytics (CBE)

College of Business and Economics

Signature Date

1. Member:

Title:

Department of … College of …

Signature Date

1. Member:

Title:

Department of … College of …

Signature Date

1. Member :

Title:

Department of …

Institution:

Signature Date

This Master Thesis is accepted by:

Dean of the College of Business and Economics:

Signature

Date

Dean of the College of Graduate Studies:

Signature

Date

# Abstract

This report presents a comprehensive methodology for analyzing public procurement data globally, focusing on efficiency, transparency, and corruption risk assessment. Specifically, the report explores how large-scale datasets, such as the Global Public Procurement Dataset (GPPD), can uncover critical insights into procurement practices across 42 countries from 2006 to 2021. The methodology applied includes advanced data analytics techniques such as trend analysis, risk assessment, and efficiency evaluation, supported by visualizations and statistical summaries.

The analysis provides a robust framework for understanding how procurement processes operate, identifying potential inefficiencies, and evaluating corruption risks. By leveraging this dataset, the report outlines strategies for enhancing transparency and improving procurement outcomes. Additionally, the methodology allows for the development of advanced risk indicators, which can guide future research and policymaking to promote accountability and efficient resource allocation in public procurement. The proposed framework not only highlights key procurement trends but also sets the stage for continuous improvement in governance through data-driven insights.

**Keywords**: Public Procurement, Corruption Risk, Data Analytics, Transparency, Global Public Procurement Dataset, Governance

# Title and Abstract (in Arabic)

**تحليل قاعدة بيانات المشتريات العامة العالمية (GPPD)**

يقدم هذا التقرير منهجية شاملة لتحليل بيانات المشتريات العامة على المستوى العالمي، مع التركيز على الكفاءة والشفافية وتقييم مخاطر الفساد. يسلط التقرير الضوء على كيفية استخدام قواعد البيانات الضخمة مثل قاعدة بيانات المشتريات العامة العالمية (GPPD) لاستخلاص رؤى مهمة حول ممارسات المشتريات في 42 دولة خلال الفترة من 2006 إلى 2021.

تشمل المنهجية المستخدمة تقنيات متقدمة لتحليل البيانات مثل تحليل الاتجاهات، تقييم المخاطر، وتقييم الكفاءة، مدعومة بتصورات بيانية وملخصات إحصائية. يوفر التحليل إطارًا قويًا لفهم كيفية عمل عمليات المشتريات، وتحديد مواطن القصور المحتملة، وتقييم مخاطر الفساد. من خلال الاستفادة من هذه القاعدة البيانية، يحدد التقرير استراتيجيات لتعزيز الشفافية وتحسين نتائج المشتريات. بالإضافة إلى ذلك، تسمح المنهجية المطروحة بتطوير مؤشرات مخاطر متقدمة، والتي يمكن أن توجه البحث والسياسات المستقبلية لتعزيز المساءلة وتحسين تخصيص الموارد في المشتريات العامة. الإطار المقترح لا يسلط الضوء فقط على الاتجاهات الرئيسية في المشتريات، ولكنه يمهد أيضًا الطريق لتحسين مستمر في الحوكمة من خلال الرؤى المستندة إلى البيانات.

الكلمات المفتاحية\*\*: المشتريات العامة، مخاطر الفساد، تحليل البيانات، الشفافية، قاعدة بيانات المشتريات العامة العالمية، الحوكمة.

# Acknowledgements

My thanks go to my committee for their guidance, support, and assistance throughout my preparation of this thesis, especially my advisor Dr. Gurdal Ertek, for his constant help and advice. He gave me important suggestions and helped me understand the material better. I am grateful for his wisdom and guidance.

Special thanks go Dr. Ananth Chiravuri, our program coordinator, for always guiding us and making our program educational and easy to follow.

My thanks also go to my professors, who have influenced my project and my future. They explained difficult subjects in a way that was easy to understand, and I am lucky to have learned from them. I appreciate all the time they spent helping us learn.

Lastly, I want to say a big thank you to my family and friends for their endless support and constant encouragement. Thank you for always believing in me.

# Dedication

*To my beloved parents and family*

Contents

[Abstract 2](#_Toc180056034)

[1. Introduction 2](#_Toc180056035)

[2. Literature Review 3](#_Toc180056036)

[3. Background 3](#_Toc180056037)

[4. Data Description 4](#_Toc180056038)

[4.1 Introducing the Data Attributes 5](#_Toc180056039)

[4.2 Example for data understanding 8](#_Toc180056040)

[5. Methodology 10](#_Toc180056041)

[**5.1 Data Collection** 10](#_Toc180056042)

[**5.2 Data Standardization** 10](#_Toc180056043)

[**5.3 Corruption Risk Indicators** 10](#_Toc180056044)

[**5.4 Data Quality and Updates** 10](#_Toc180056045)

[6. Data Analysis 10](#_Toc180056046)

[**6.1 Prompt engineering** 11](#_Toc180056047)

[6.1.1 Zero Shot 12](#_Toc180056048)

[6.1.2 Mempromt 13](#_Toc180056049)

[6.1.3 Chain of Thought 13](#_Toc180056050)

[6.1.4 Self Ask 13](#_Toc180056051)

[6.1.5 Prompt Chaining 14](#_Toc180056052)

[6.1.6 ACT 14](#_Toc180056053)

[6.1.7 ReAct 14](#_Toc180056054)

[6.1.8 Few Shot 14](#_Toc180056055)

[6.1.9 Inception 14](#_Toc180056056)

[6.1.10 Memetic Proxy 15](#_Toc180056057)

[6.1.11 Self-consistency 15](#_Toc180056058)

[6.1. 12 Action Plan Generation and Execute Technique 15](#_Toc180056059)

[**6.2 Preparing the Data** 16](#_Toc180056060)

[6.3 Analysis Techniques 16](#_Toc180056061)

[6.4 Advanced Techniques 17](#_Toc180056062)

[6.5 Reporting and Visualization 18](#_Toc180056063)

[6.6 Findings and Insights 19](#_Toc180056064)

[6.7 Dashboard 24](#_Toc180056065)

[7. Public Procurement Analyst GPT 25](#_Toc180056066)

[8. Conclusion 26](#_Toc180056067)

[9. Statistics 27](#_Toc180056068)

[References 27](#_Toc180056069)

[Appendices 28](#_Toc180056070)

[**Filters** 31](#_Toc180056071)

[**Corruption Risk Indicators** 31](#_Toc180056072)

**Figures:**

4.1.Countries covered in the Government Transparency Institute Global Public Procurement Dataset….…………………………………………………………………………………………………………………………………………………4

* + 1. Annual total number of contracts by continent across time…………………………………………………….....8
  1. Prompt Engineering Programming Patterns……………………………………………………………………………………11

## Abstract

One-third of total government spending across the globe goes to public procurement, amounting to about 10 trillion dollars a year. Despite its vast size and crucial importance for economic and political developments, there is a lack of globally comparable data on contract awards and tenders. To fill this gap, the Global Public Procurement Dataset (GPPD) was introduced. Using web scraping methods, official public procurement data on over 72 million contracts from 42 countries between 2006 and 2021 were collected. The data was standardized to fit a common format, and several corruption risk indicators were added to enable objective assessment of risks. This dataset offers rich, contract-level information crucial for understanding and analyzing public spending efficiency, corruption risks, and market dynamics.

1. **Introduction**

* **Research and Policy Analysis:** Researchers and policy analysts studying public spending can benefit from this global, standardized, micro-level dataset. It provides rich, contract-level information on where and how governments spend public funds, accounting for about one-third of general government spending in the countries covered by the data.
* **Monitoring and Analysis:** Academics, governments, and control bodies (e.g., auditors) can use the data to monitor and analyze public procurement across a wide range of countries, including tracking corruption risks.
* **Data Integration:** Government contracts data can be linked to other datasets, increasing its value. For example, it can be connected to company registry data or politicians’ asset declarations to gain comprehensive insights into public spending quality and good governance.
* **Enhanced Insights:** This dataset adds value to existing macro-level datasets on public spending by providing comprehensive contract-level information. Micro-level data on the process and outputs of public procurement spending help analyze market dynamics and spending efficiency.

## Literature Review

Public procurement accounts for a significant portion of government spending, estimated at around 10 trillion USD annually worldwide (OECD, 2016). Despite its vast size and importance, there has been a lack of globally comparable data on public procurement contracts and tenders. The Global Public Procurement Dataset (GPPD) aims to fill this gap by collecting and harmonizing public procurement data from various countries.The GPPD builds upon previous efforts to monitor and analyze public procurement. The European Commission has emphasized the importance of public procurement for the single market and has provided guidance on best practices (European Commission, 2018). The World Bank has also developed guidelines for procurement in its investment projects (World Bank, 2020). However, these initiatives have typically focused on specific regions or sectors.Researchers have long recognized the need for comprehensive, micro-level data on public procurement to study market dynamics, efficiency, and corruption risks. Bosio et al. (2020) highlighted the importance of public procurement data for understanding the relationship between law and practice. Fazekas and Tóth (2016) used contract-level data to assess corruption risks in Hungarian public procurement. The GPPD dataset provides a unique opportunity to conduct cross-country comparisons and identify global trends in public procurement. By standardizing data from various sources and adding corruption risk indicators, the dataset enables researchers and policymakers to assess the quality of public spending and identify areas for improvement. The literature emphasizes the need for transparency and accountability in public procurement to mitigate corruption risks (OECD, 2016; European Commission, 2018). The GPPD dataset supports these efforts by providing a tool for monitoring and analyzing public procurement across a wide range of countries. In conclusion, the Global Public Procurement Dataset fills an important gap in the literature by providing a comprehensive, standardized dataset on public procurement contracts. The dataset enables researchers and policymakers to study market dynamics, efficiency, and corruption risks in public spending, ultimately contributing to more transparent and accountable governance.

**3. Background**

Public procurement is a crucial area of public spending, amounting to about one-third of general government spending worldwide, worth around 9.5 trillion USD annually. Such large amounts are accompanied by high public interest as key infrastructure and services depend on government contracts. However, public procurement faces high corruption risks due to its complexity and high degrees of discretion.

Many countries publish large amounts of micro-level information on public procurement. Unfortunately, this information is often hard to use due to inconsistent formats and fragmentation across different websites and legal regimes. While a few datasets allow monitoring public procurement performance, they typically cover only one country or sector, lacking the scale and scope of the GPPD.

**4. Data Description**

Public procurement procedures are highly regulated and structured processes. A typical open public procurement tender starts with a call for tenders or request for quotations. Potential suppliers submit their bids during the advertisement period, which are then evaluated by a tender evaluation committee. After the contract award decision, a notice is published, and the contract is concluded between the buyer and the supplier. The procurement process completes with the delivery according to the contract or its termination.

The GPPD includes harmonized public procurement contracts from 42 countries, covering data mainly from 2006 to 2021. The dataset, published by the Government Transparency Institute, includes over 72 million contracts involving around 1.8 million buyers and more than 10 million suppliers. It represents a total contract value of more than USD 16.8 trillion, which is about 1.1% of global GDP annually.

Public procurement procedures are highly regulated and tightly structured. A typical open public procurement tender begins with a call for tenders or request for quotations. During this period, potential suppliers submit their bids, which are then evaluated by a tender evaluation committee. After the contract award decision is made and published, the contract is concluded between the buyer and the supplier, followed by contract implementation and completion.

Our dataset includes harmonized public procurement contracts from 42 countries, covering data mainly from 2006 to 2021. The dataset includes over 72 million contracts involving approximately 1.8 million buyers and more than 10 million suppliers. The total contract value exceeds USD 16.8 trillion, representing about 1.1% of global GDP annually.

The data is available for countries in dark blue in the figure below

A map of the world

Description automatically generated

Fig. 1. Countries covered in the Government Transparency Institute Global Public Procurement Dataset.

The GPPD is a comprehensive, unfiltered dataset that includes harmonized contract data and procurement notices that failed or were canceled. Each contract in the dataset represents a commitment to public spending. The data covers various types of procurement, including services, works, and supplies. Services contracts, such as medical treatments and education services, account for USD 6.48 trillion. Works contracts, like construction and refurbishment, total USD 3.35 trillion. Supplies contracts, encompassing goods like cars and office supplies, represent USD 2.23 trillion.

The dataset is continuously updated by the Government Transparency Institute to improve data quality and include recent records. Users can track updates using the variable “persistent\_id”, which remains consistent across dataset releases. This enables researchers to identify updated records and maintain the integrity of their analyses. The dataset can be downloaded by country from the Mendeley repository, providing researchers with a valuable resource for studying public procurement on a global scale.

### 4.1 Introducing the Data Attributes

The Global Public Procurement Dataset (GPPD) includes various attributes that provide detailed information about each procurement process. Below are key data attributes from the dataset:

* **persistent\_id**: A unique identifier for each tender, created by hashing the URL of the tender's first publication.
* **tender\_id**: An internal identifier assigned during data processing.
* **tender\_title**: The title of the tender.
* **tender\_proceduretype**: The type of procedure used for the tender, standardized to common categories (e.g., Open, Restricted, Negotiated).
* **tender\_nationalproceduretype**: The original procedure type as published, including jurisdiction-specific terms.
* **tender\_isawarded**: Indicates if the tender was awarded (true or false).
* **tender\_supplytype**: The type of purchase: supplies, services, or public works.
* **tender\_biddeadline**: The final deadline for bid submissions.
* **tender\_isjointprocurement**: Indicates if the tender is a joint procurement involving multiple public bodies.
* **tender\_lotscount**: The number of lots (subsections) in the tender.
* **tender\_recordedbidscount**: The number of unique bids recorded for the tender.
* **tender\_isframeworkagreement**: Indicates if the tender is a framework agreement (a long-term agreement for multiple purchases).
* **tender\_isdps**: Indicates if the tender is a Dynamic Purchasing System (a flexible way to manage procurement).
* **tender\_contractsignaturedate**: The date the contract was signed.
* **tender\_cpvs**: List of product codes for the items purchased, based on Common Procurement Vocabulary (CPV) codes.
* **tender\_maincpv**: The main product code for the tender, based on CPV codes.
* **tender\_iseufunded**: Indicates if the tender is funded by the European Union.
* **tender\_selectionmethod**: The criteria for selecting the winning supplier: lowest price or most economically advantageous tender (MEAT).
* **tender\_awardcriteria\_count**: The number of criteria used to evaluate the bids.
* **tender\_cancellationdate**: The date the tender was canceled.
* **cancellation\_reason**: The reason for the tender or contract cancellation.
* **tender\_awarddecisiondate**: The date the award decision was made.
* **tender\_estimatedprice**: The estimated price of the tender.
* **tender\_finalprice**: The final price of the tender.
* **lot\_estimatedprice**: The estimated price for each lot within the tender.
* **bid\_price**: The price of a specific bid.
* **tender\_corrections\_count**: The number of corrections made to the tender.
* **lot\_row\_nr**: A unique identifier for each lot within a tender.
* **lot\_title**: The title of the lot.
* **lot\_status**: Indicates if the lot was awarded.
* **lot\_bidscount**: The total number of bids submitted for the lot.
* **lot\_validbidscount**: The total number of valid bids for the lot.
* **lot\_electronicbidscount**: The number of bids submitted electronically for the lot.
* **lot\_smebidscount**: The number of bids submitted by Small and Medium-sized Enterprises (SMEs) for the lot.
* **lot\_updateddurationdays**: The latest duration of the lot or contract in days.
* **buyer\_id**: The main identifier of the buyer from the source documents.
* **buyer\_masterid**: A unique identifier of the buyer assigned during data processing.
* **buyer\_name**: The name of the buyer.
* **buyer\_nuts**: The regional code of the buyer, based on the Nomenclature of Territorial Units for Statistics (NUTS).
* **buyer\_city**: The city of the buyer.
* **buyer\_country**: The country of the buyer.
* **buyer\_mainactivities**: The main activities of the buyer (e.g., health, education, public services).
* **buyer\_buyertype**: The type of buyer (e.g., national authority, regional authority, public body).
* **buyer\_postcode**: The postcode of the buyer.
* **buyer\_nuts\_1**: The first-level NUTS code of the buyer's location.
* **buyer\_nuts\_2**: The second-level NUTS code of the buyer's location.
* **buyer\_nuts\_3**: The third-level NUTS code of the buyer's location.
* **buyer\_street**: The street address of the buyer.
* **buyer\_url**: The website of the buyer.
* **buyer\_email**: The email address of the buyer.
* **buyer\_phone**: The phone number of the buyer.
* **buyer\_contactName**: The contact person's name at the buyer's organization.
* **buyer\_extra\_source\_id**: Other identifiers of the buyer from source documents.
* **buyer\_sourceid\_type**: The type of other identifiers for the buyer.
* **bidder\_id**: The main identifier of the bidder company from the source documents.
* **bidder\_masterid**: A unique identifier of the bidder company assigned during data processing.
* **bidder\_name**: The name of the bidder company.
* **bidder\_nuts**: The regional code of the bidder company, based on NUTS.
* **bidder\_city**: The city of the bidder company.
* **bidder\_country**: The country of the bidder company.
* **bidder\_postcode**: The postcode of the bidder company.
* **bidder\_street**: The street address of the bidder company.
* **bidder\_email**: The email address of the bidder company.
* **bidder\_phone**: The phone number of the bidder company.
* **bidder\_extra\_source\_id**: Other identifiers of the bidder company from source documents.
* **bidder\_sourceid\_type**: The type of other identifiers for the bidder company.
* **bidder\_url**: The website of the bidder company.
* **bidder\_contactName**: The contact person's name at the bidder company.
* **bidder\_nuts\_3**: The third-level NUTS code of the bidder company's location.
* **bidder\_nuts\_2**: The second-level NUTS code of the bidder company's location.
* **bidder\_nuts\_1**: The first-level NUTS code of the bidder company's location.
* **bid\_iswinning**: Indicates if the bid was a winning bid (true or false).
* **bid\_issubcontracted**: Indicates if part of the contract will be subcontracted.
* **bid\_subcontractedproportion**: The portion of the contract expected to be subcontracted.
* **bid\_isconsortium**: Indicates if the bid is submitted by a consortium.
* **source**: The source of the tender data.
* **tender\_publications\_lastcontractawardurl**: The URL of the last contract award announcement.
* **tender\_publications\_firstdcontractawarddate**: The date of the first contract award announcement.
* **notice\_url**: The URL of the last call for tenders or contract notice publication.
* **tender\_publications\_firstcallfortenderdate**: The date of the first call for tender announcement.
* **tender\_year**: The year of the tender.
* **tender\_addressofimplementation\_nuts**: The regional code of the tender implementation location based on NUTS.
* **tender\_description\_length**: The length of the tender description in characters.
* **lot\_description\_length**: The length of the lot description in characters.
* **tender\_personalrequirements\_length**: The length of the personal requirements for participation in characters.
* **tender\_technicalrequirements\_length**: The length of the technical requirements for participation in characters.
* **tender\_economicrequirements\_length**: The length of the economic requirements for participation in characters.
* **currency**: The currency of the prices.
* **tender\_digiwhist\_price**: The estimated final price of the tender using various methods if the actual final price is not available.
* **bid\_digiwhist\_price**: The estimated contract price using the bid price or lot estimated price if the bid price is missing.
* **lot\_id**: A unique identifier for a lot assigned during data processing.
* **bid\_id**: A unique identifier for a bid assigned during data processing.
* **bid\_priceUsd**: The bid price converted to International USD.
* **lot\_estimatedpriceUsd**: The estimated lot price converted to International USD.
* **tender\_estimatedpriceUsd**: The estimated tender price converted to International USD.
* **tender\_finalpriceUsd**: The final tender price converted to International USD.

All those attributes have so many meaningful information that we can use and extract helpful informaton from them, for example, the annual total number of contracts by continent across time.

### 4.2 Example for data understanding

The dataset is diverse in terms of annual country coverage, encompassing 33 countries in Europe, 4 in South America, 2 in North America, 2 in Asia, and 1 in Africa. The figure below. shows the annual distribution of awarded contracts per continent.

Since 2015, there has been a consistent rise in the number of published contracts in Europe and North America, peaking at approximately 5 million contracts per year in 2019–2020. However, the drop in the European contract count in 2021 is attributed to changes in publication practices in some countries, such as Italy, which switched to a new open data publication format. This new format will need to be retrospectively incorporated into the database during future updates. South America has also seen a steady increase in the number of contracts, reaching around 1.5 million contracts by 2021. Also, there has been a rise in the number of published contracts in Asia and Africa, although less so than other continents.

This visualization highlights the trends in public procurement across different continents over the years, illustrating the growing transparency and availability of procurement data globally.

A graph with different colored bars

Description automatically generated

Fig. 2. Annual total number of contracts by continent across time.

As a conclusion for this part, The Global Public Procurement Dataset (GPPD) is a comprehensive resource that provides detailed insights into public procurement processes worldwide. Covering over 72 million contracts from 42 countries between 2006 and 2021, the dataset standardizes diverse procurement data into a unified format, enabling comparative analysis and monitoring of procurement activities.

Key features of the dataset include:

* **Wide Coverage**: The dataset spans 33 countries in Europe, 4 in South America, 2 in North America, 2 in Asia, and 1 in Africa.
* **Rich Information**: It includes detailed attributes for each tender, such as procedure types, supply types, bid deadlines, contract values, and corruption risk indicators.
* **Temporal Trends**: The data shows consistent growth in the number of published contracts, particularly in Europe and North America, with noticeable increases in South America, Asia, and Africa.
* **Data Quality and Updates**: The dataset is regularly updated to improve data quality and incorporate recent records.

By providing harmonized contract data and including procurement notices that were canceled or failed, the GPPD offers a robust foundation for analyzing public spending efficiency, corruption risks, and market dynamics. The dataset’s extensive coverage and detailed information make it a valuable tool for researchers, policymakers, and auditors aiming to enhance transparency and efficiency in public procurement.

Next, we will explore the Experimental Design, Materials, and Methods used to compile and standardize this extensive

**5. Methodology**

**5.1 Data Collection**

The dataset creation consists of three main stages: data collection, data standardization, and data validation.

First, the data collection stage begins by identifying official government publication portals that provide public procurement information. These sources may include online tender journals or structured databases offering procurement data. An automated web crawler is then developed to scrape data from these portals, with some countries having multiple sources. The collected data includes HTML, XML, JSON, and CSV files from official government sources. Data collection is limited to countries that publish semi-structured online data on their public procurement procedures, adhering to standardized formats defined by legal regulations.

This thorough data collection process ensures that a wide range of procurement information is gathered from various countries, forming the foundation for the subsequent stages of data processing and analysis.

**5.2 Data Standardization**

The data standardization process involves several steps to improve data accuracy and usability. Missing tender years are imputed using related publication years. Procedure types are harmonized into common categories, and erroneous organization names are cleaned. NUTS codes are restructured to standardize regional information for European countries. A "filter\_framework" flag identifies primary contracts from framework agreements. Prices are adjusted using PPP conversion factors from the World Bank to enable cross-country comparisons. These steps ensure a clean, standardized, and comprehensive dataset ready for meaningful analysis.

**5.3 Corruption Risk Indicators**

Several corruption risk indicators were added to the dataset, along with a composite corruption risk index for each contract. These indicators enable an objective assessment of corruption risks and facilitate comparisons across time, organizations, or countries.

**5.4 Data Quality and Updates**

The research team routinely updates the dataset to improve data quality and add recent data. Updated records are identified using the variable **persistent\_id**, which remains consistent across dataset updates. If records with the same **persistent\_id** differ in different dataset releases, the record has been updated.

**6. Data Analysis**

In the data analysis phase, we will leverage advanced prompt engineering techniques in combination with GPT (Generative Pre-trained Transformer) to analyze and update information from the Global Public Procurement Dataset (GPPD). This approach will enable us to extract meaningful insights, ensure data accuracy, and provide timely updates. Prompt engineering involves designing specific queries and commands that guide GPT to process and analyze the dataset effectively. By creating tailored prompts, we can instruct GPT to perform complex data analysis tasks, such as identifying trends, detecting anomalies, and assessing corruption risks.

But let’s start by defining what is the prompt engineering

**6.1 Prompt engineering**

Prompt engineering means creating and organizing the questions or commands (prompts) given to an AI (Artificial Intelligence) to get useful and relevant answers. How well the user designs these prompts affects the quality and relevance of the AI's responses. Good prompt engineering involves knowing what the AI can and cannot do, and using this understanding to make prompts that guide the AI to produce the best results.

Prompt Engineering is important and powerful tool for increasing the Precision and efficiency of a model, the versatility, and the control, here we talk about how user can control the type of response generated.

There are so many prompt techniques, here we will focus on text-to-text technique.

And we will give prompts for ' Data Exploration and Interpretation', 'Data Cleaning and Preprocessing’,’ Statistical Analysis and Hypothesis Testing’,’ Data Visualization and Reporting', 'Coding and Scripting Assistance’ and ‘Sentiment Analysis’.

We will apply 12 prompts engineering technique in the figure below.

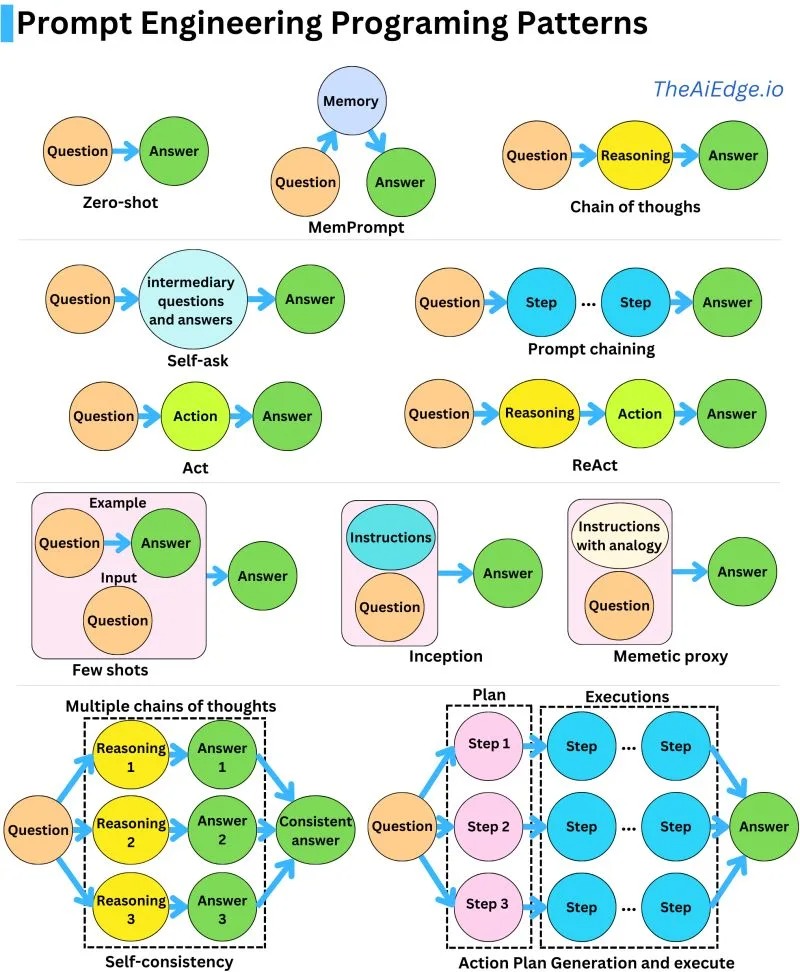


Fig.6.1. Prompt Engineering Programming Patterns

### 6.1.1 Zero Shot

Zero-shot prompting means the user asks the AI to do something without giving AI any examples to follow. The user just tells the AI what he wants it to do directly.

Imagine a chef has a new recipe book, and he asks his employee to bake a cake from it without showing him a picture or an example of how the cake should look. He must follow the instructions without any help

Similarly, with zero-shot prompting, the user gives the AI a task, and the AI must figure out how to do it without examples.

#### Example

*Data Exploration and Interpretation:*

"What are the top 10 most common tender titles in the dataset?"

*Data Cleaning and Preprocessing:*

"Identify columns with missing values and count the number of missing entries in each column."

*Statistical Analysis and Hypothesis Testing:*

"Calculate the correlation between tender values and the number of bids."

*Data Visualization and Reporting:*

"Create a bar chart showing the frequency of the top 5 tender titles."

*Coding and Scripting Assistance:*

"Write a script to load the dataset and display the first few rows using python language."

*Sentiment Analysis:*

"Analyze the distribution of sentiments (positive, neutral, negative) in the tender titles."

### 6.1.2 Mempromt

MemPrompt uses memory from past interactions or stored knowledge to help answer current questions or tasks. By remembering what was done before, the AI gives better and more relevant responses. for example, if someone asks his friend to recall and summarize a conversation he had last week, he uses his memory of that discussion to give him an answer.

You will finds the examples in the analysis file

### 6.1.3 Chain of Thought

Chain-of-thought (CoT) prompting helps the AI think through complex problems by breaking them down into smaller steps. The user can combine this with few-shot prompting, where he gives a few examples, to get better results for tasks that need reasoning before answering.

Imagine a teacher who is teaching his students to solve a math problem. First, this teacher shows them a few examples of how to solve similar problems step by step. Then, he guides them through each step of a new problem, helping them understand the reasoning behind each part.

Similarly, chain-of-thought prompting enables complex reasoning capabilities through intermediate reasoning steps.

### 6.1.4 Self Ask

Self-ask prompting allows AI to show its thinking process by breaking down complex questions into smaller follow-up questions. AI explicitly shows its reasoning and knows when it has reached the last answer. It seamlessly moved from these intermediate steps to the last answer.

Imagine a father who teaches his child to solve a puzzle. Instead of just solving it for him, he guides him to ask themselves questions like, "What pieces fit together?" and "What’s the next piece I need?" He breaks down the problem, step by step, until he solves the puzzle

### 6.1.5 Prompt Chaining

Prompt chaining is useful for handling complex tasks that an AI might find difficult if given all at once. In prompt chaining, each prompt builds on the previous one, performing transformations or additional steps on the generated responses until the last answer is reached. This method is often used to answer questions about a large text document, imagine a supervisor is helping his student to author a detailed report. Instead of asking him to write it all at once, this supervisor first asks him to outline the main sections. Then, he asks him to write each section one by one, and finally, he asks him to review and edit the entire report.

### 6.1.6 ACT

Performing actions based on direct instructions to achieve a specific goal, for example, following a recipe involves measuring ingredients, mixing the ingredients, and baking to achieve the desired dish.

### 6.1.7 ReAct

ReAct (Reasoning and Acting) is a method where a user thinks about a problem, takes action to solve it, and then repeats this process. This back-and-forth approach helps the user to make better decisions and adjustments based on what he learns each time.it helps him to solve problems and make decisions by thinking, acting, and learning in cycles. For example, when troubleshooting a device, technician reasons about probable causes of the issue, takes action to fix it, and then evaluates the results, repeating this cycle, as necessary.

### 6.1.8 Few Shot

Few-shot prompting is when the AI has a memory to help it understand how to complete a task. These memories guide the AI and improve its performance on similar tasks, AI can be used as a technique to enable in-context learning.

We will adopt the same example as zero shot technique but, in this case, Imagine teaching someone to bake a cake by showing him how to make a few distinct types of cakes first, he learns the general steps and can then makes other cakes more easily.

### 6.1.9 Inception

Using detailed instructions to guide the completion of a task. This approach ensures that complex tasks are broken down into clear, manageable steps, making it easier to achieve the desired outcome. For example, a tourist guide follows a detailed map with instructions to reach a specific destination, ensuring that the tourist makes the right turns and follows the correct path.

### 6.1.10 Memetic Proxy

The Memetic Proxy technique involves using analogies or metaphors to explain or perform tasks. This approach helps make complex concepts more understandable by relating them to familiar or simpler ideas. For example, imagine a researcher with a Large Text Document Breaks down the task: first summarize the document, then extract key data, analyze trends, and finally answer specific questions. This step-by-step approach helps tackle complex tasks efficiently.

### 6.1.11 Self-consistency

The Self-Consistency technique means doing the same task multiple times and comparing the results to make sure they match. This helps confirm that the results are reliable and accurate. For example, if a researcher is measuring the height of a plant, he must first, Measures the plant three times, Compares the three measurements, checks if any measurement is different, fix any mistakes (like using a straight measuring tape), Finally measures again to ensure all heights are the same.

### 6.1. 12 Action Plan Generation and Execute Technique

The action plan generation and execution technique involve creating a detailed plan for completing a complex task and then executing that plan step by step. This method ensures that each aspect of the task is carefully considered and systematically addressed, leading to well-organized and effective execution. For example, a party organizer Identifies goals like having fun and celebrating, then breaks down tasks like sending invites and decorating. Assigns deadlines and executes by creating a guest list and setting up decorations.

The integration of GPT with the GPPD will allow us to:

* **Automated Data Analysis**: Utilize GPT to automatically analyze procurement data, highlighting key metrics and trends.
* **Generate Insights**: Produce comprehensive reports and visualizations based on the latest data, making the information accessible and understandable.
* **Update Information**: Continuously update the dataset with new information and corrections, ensuring that our analysis reflects the most current data available.
* **Enhanced Decision Making**: Provide policymakers and researchers with up-to-date insights, enabling informed decision-making and targeted policy interventions.

By combining prompt engineering with GPT's powerful capabilities, we can enhance the analysis of the GPPD, ensuring that our findings are robust, relevant, and timely. This innovative approach will help us maximize the value of the dataset and support efforts to improve transparency and efficiency in public procurement.

**6.2 Preparing the Data**

**Data Cleaning**: The initial step involves ensuring that all data fields are correctly formatted and consistent across different countries. This includes:

* **Standardizing Date Formats**: Convert all date fields to a uniform format (e.g., DDMMYYYY) to ensure consistency.
* **Correcting Location Codes**: Ensure that all geographical data, such as NUTS codes, are accurately represented and consistent.
* **Handling Missing Values**: Impute missing values for critical fields like tender year using related information such as the contract award publication year or call for tender publication year. Remove erroneous data such as contracts with non-alphabetic organization names.

**Data Integration**: Enhance the dataset by linking the GPPD with other relevant datasets, such as:

* **Company Registries**: Connect public procurement data with company registry information to analyze supplier details and ownership structures.
* **Asset Declarations**: Integrate asset declaration data to investigate potential conflicts of interest and corruption risks among procurement officials.

### 6.3 Analysis Techniques

* **Descriptive Statistics**: Compute basic statistical measures (mean, median, standard deviation) to summarize key variables such as contract value, number of bids, and award dates. This helps in understanding the overall distribution and central tendencies of the data.
* **Trend Analysis**: Identify and analyze trends over time in public procurement spending, the number of contracts awarded, and the average contract value. This provides insights into how procurement activities and spending patterns evolve.
* **Geospatial Analysis**: Map the geographical distribution of contracts to identify regional patterns and disparities in public procurement. This involves visualizing contract data across different regions to detect hotspots and areas with significant procurement activities.
* **Risk Assessment**: Utilize the provided corruption risk indicators to evaluate and compare risk levels across different countries, sectors, and time periods. These indicators include single bidding prevalence, non-competitive tendering conditions, and other factors that signal potential corruption.
* **Efficiency Analysis**: Assess the efficiency of public spending by analyzing the relationship between estimated and final contract prices and the duration of contract implementation. This helps determine how well procurement processes are managed and whether public funds are used effectively.

### 6.4 Advanced Techniques

* **Regression Analysis**: Conduct regression analyses to explore the factors influencing contract award decisions and prices. Variables may include economic conditions, political factors, and procedural characteristics. This technique helps identify significant predictors and understand the determinants of procurement outcomes.
* **Network Analysis**: Investigate the relationships between buyers and suppliers using network analysis. This method identifies potential conflicts of interest, corruption networks, and the structure of procurement relationships. It reveals how entities are interconnected and where influence and dependencies lie.
* **Machine Learning**: Develop and apply machine learning models to predict high-risk contracts and flag potential cases of fraud or inefficiency. These models can use historical data to learn patterns associated with corruption and inefficiency, providing a proactive tool for risk management and fraud detection.

The data analysis phase of the Global Public Procurement Dataset (GPPD) involves thorough data cleaning, integration, and the application of both basic and advanced analytical techniques. By standardizing formats, correcting errors, and integrating supplementary datasets, we ensure the dataset's accuracy and comprehensiveness. Descriptive statistics, trend analysis, geospatial analysis, risk assessment, and efficiency analysis provide a solid foundation for understanding procurement patterns and identifying potential issues.

Advanced techniques such as regression analysis, network analysis, and machine learning enable deeper insights into the factors influencing procurement outcomes, the relationships between procurement entities, and the detection of high-risk contracts. These methods collectively support robust analysis and informed decision-making.

Incorporating GPT and prompt engineering further enhances our ability to extract meaningful information from the GPPD. Prompt engineering involves crafting specific queries and commands that guide GPT to process and analyze the dataset effectively.

By leveraging GPT and prompt engineering, we can maximize the value of the GPPD, providing a powerful tool for analyzing public procurement data, improving transparency, and enhancing the efficiency and accountability of public spending. This innovative approach ensures that our findings are robust, relevant, and actionable, ultimately contributing to better governance and public trust.

### 6.5 Reporting and Visualization

In this part, we will utilize ChatGPT prompts integrated with PowerBI to create various visualization tools, enhancing the analysis and presentation of the Global Public Procurement Dataset (GPPD). This approach will enable us to present complex data in an accessible and interactive manner, facilitating better understanding and decision-making.

**Dashboards**:

We will create interactive dashboards using PowerBI to visualize key metrics and trends in public procurement. These dashboards will provide a real-time, dynamic view of the data, allowing users to explore different aspects of procurement activities, such as contract values, the number of bids, and geographical distributions. Key features of the dashboards will include:

* **Trend Analysis**: Visualize trends over time in procurement spending and contract awards.
* **Geospatial Analysis**: Map the distribution of contracts to identify regional patterns.
* **Risk Indicators**: Highlight corruption risk indicators across different countries and sectors.

**Reports**:

Detailed reports will be generated to summarize findings from the dataset. These reports will provide actionable insights for policymakers and stakeholders, covering areas such as procurement efficiency, corruption risks, and spending patterns. The reports will be structured to include:

* **Descriptive Statistics**: Summaries of key variables like contract value, number of bids, and award dates.
* **Efficiency Analysis**: Evaluations of public spending efficiency by comparing estimated and final contract prices.
* **Corruption Risk Assessments**: Analyses using the provided corruption risk indicators to compare risk levels across different regions and sectors.

**Infographics**:

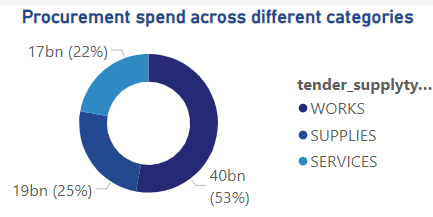
To communicate complex data and findings in an easily understandable format, we will design infographics. These visual representations will distill essential insights from the dataset into clear and engaging graphics, making the information accessible to a broader audience. Infographics will include:

* **Summary of Key Metrics**: Visual summaries of important data points such as total contract value and procurement trends.
* **Risk Indicator Highlights**: Infographics that emphasize key corruption risk indicators and their implications.
* **Comparative Analysis**: Visual comparisons of procurement activities and risks across different countries.

By integrating ChatGPT prompts with PowerBI, we will ensure that our visualizations are not only informative but also interactive and user-friendly. This combination of advanced AI capabilities with powerful visualization tools will enhance our ability to analyze and present the GPPD, making it a valuable resource for researchers, policymakers, and the public.

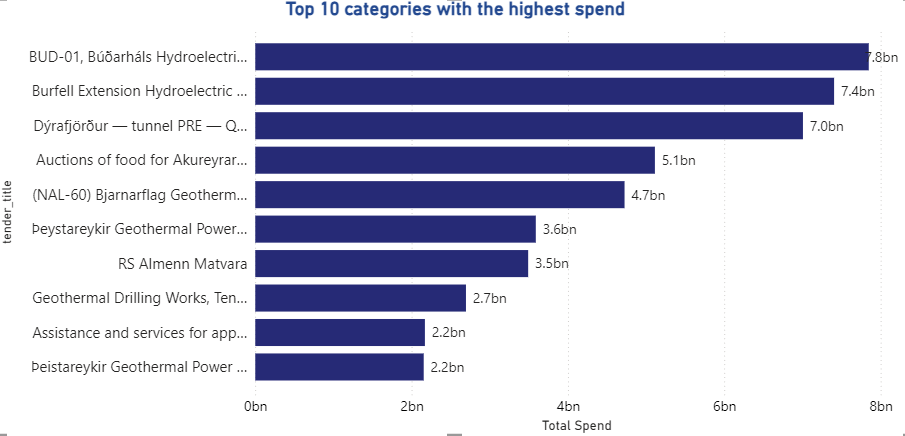
The integration of these visualization techniques will provide a comprehensive and visually engaging way to explore the GPPD, supporting informed decision-making and promoting transparency in public procurement.

### 6.6 Findings and Insights



The analysis reveals that **WORKS** is the dominant category in procurement spending, constituting approximately **52.9%** of the total budget, equivalent to **39.9 billion**. This highlights a strategic focus on infrastructure and construction projects. **SUPPLIES** follows as the second largest category with **24.9%** of the total spend, amounting to **18.8 billion**, indicating significant investment in essential goods for public services. The **SERVICES** category, although the smallest among the major categories, still represents **22.2%** of the total spend with **16.8 billion**, showing its importance in supporting operations.

The high allocation towards **WORKS** suggests an opportunity for efficiency reviews to optimize expenditures in this area. The balanced distribution of resources across **SUPPLIES** and **SERVICES** supports a diversified investment strategy, but it is essential to ensure these funds are aligned with strategic goals and public needs.

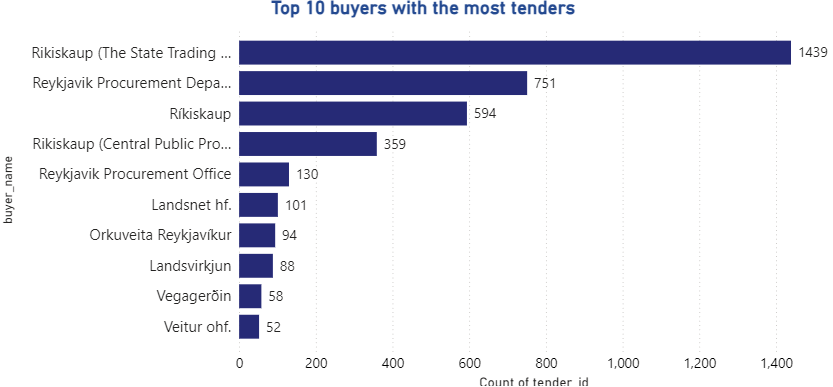


**Findings**:

1. **Top Tender:**
   * The tender titled **"BUD-01, Búðarháls Hydroelectric Project, no. 20015, Civil Works"** has the highest total spend, amounting to **7.85 billion**. This indicates a major investment focus on infrastructure development, specifically in hydroelectric projects.
2. **Subsequent High-Spending Tenders:**
   * The **"Burfell Extension Hydroelectric Project, Civil Works"** follows closely with a spend of **7.40 billion**, and **"Dýrafjörður — tunnel PRE — Qualification"** is next with **7.01 billion**, indicating significant budget allocations toward energy and infrastructure.
   * Other tenders such as **"Auctions of food for Akureyrarbæ"** (5.11 billion) and **"(NAL-60) Bjarnarflag Geothermal Power Plant and Þeystareykir Geothermal Plant - Engineering Consultancy Services"** (4.72 billion) demonstrate the diversity of high-value projects beyond just civil works.
3. **Diverse Allocation:**
   * The remaining tenders, including **"Þeystareykir Geothermal Power Plant, Buildings"** (3.59 billion) and **"RS Almenn Matvara"** (3.49 billion), show investments spanning across various sectors like food supply and geothermal energy.

**Insights**:

* The concentration of large expenditures in hydroelectric and geothermal projects indicates a strategic focus on sustainable energy and infrastructure development.
* The allocation for food supply and social assistance services reflects a commitment to supporting essential community needs alongside infrastructure projects.



**Findings and Insights:**

1. **Leading Buyer:**

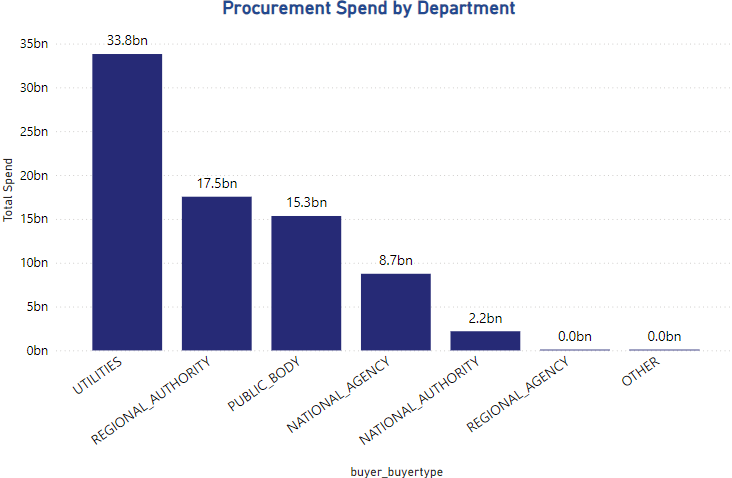
* The highest number of tenders was initiated by **Rikiskaup (The State Trading Centre)** with **1,439** tenders. This indicates their central role in procurement activities, likely managing a broad range of contracts for various government needs.

1. **Other Major Buyers:**

* The **Reykjavik Procurement Department** follows with **751** tenders, and **Ríkiskaup** has **594** tenders. These entities are also highly active, supporting a wide array of public procurement initiatives.

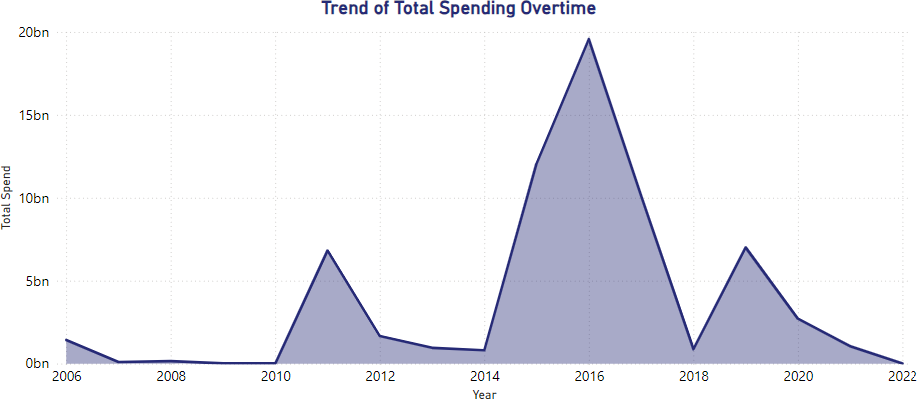
1. **Diverse Buyer Involvement:**

* The remaining buyers, such as **Landsnet hf.** with **101** tenders and **Orkuveita Reykjavíkur** with **94**, show that beyond the primary agencies, several specialized entities also play significant roles in procurement.



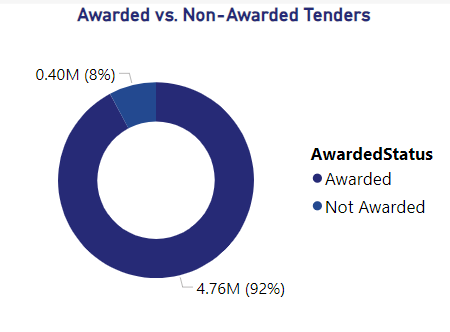
The department type **"UTILITIES"** leads with a total spend of **33.81 billion**, indicating a significant focus on utility-related projects and services. This is followed by **"REGIONAL AUTHORITY"** and **"PUBLIC BODY"** with spends of **17.54 billion** and **15.34 billion** respectively, highlighting their prominent roles in managing public funds. **"NATIONAL AGENCY"** and **"NATIONAL AUTHORITY"** also feature significant expenditures, while smaller spends are noted for regional and European agencies.

The distribution reflects a diverse allocation approach, prioritizing essential services and infrastructure managed by utilities and regional authorities. Efficiency audits for high-spend departments like **"UTILITIES"** and **"REGIONAL AUTHORITY"** could optimize spending further.



The spending trend over time shows notable variations:

* In **2006**, the spend was **1.4 billion**, followed by a drop to **97.8 million** in **2007** and **148.9 million** in **2008**.
* A significant increase occurred in **2011**, reaching **6.8 billion**, indicating major projects or initiatives.
* **2015** and **2016** marked the highest peaks with **12.1 billion** and **19.6 billion** respectively, reflecting extensive procurement activities, likely tied to large-scale investments.
* Following these peaks, there was a decline, with **2018** seeing **866.3 million** and **2021** at **1.0 billion**.
* The most recent data point in **2022** shows a sharp drop to **1.5 million**, which might indicate reduced activity or incomplete data for that year.



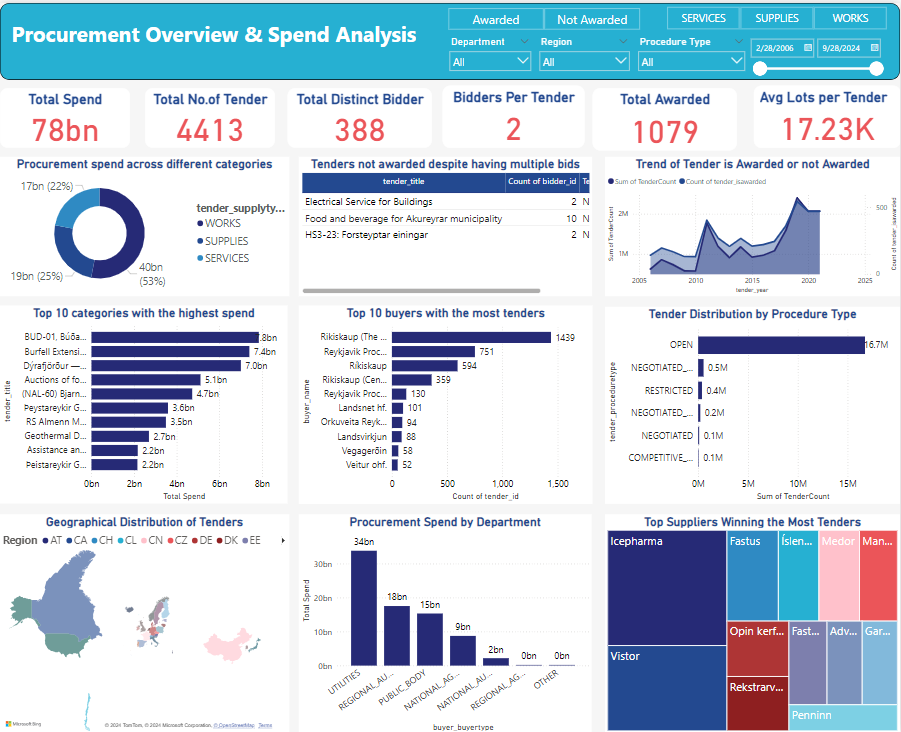
The data shows that **92.2%** of tenders (1,079 in total) were successfully awarded, demonstrating a robust procurement process with a high rate of contract fulfillment. The **7.8%** of tenders (91 in total) that were not awarded, however, highlight potential inefficiencies or challenges.

**Insights**:

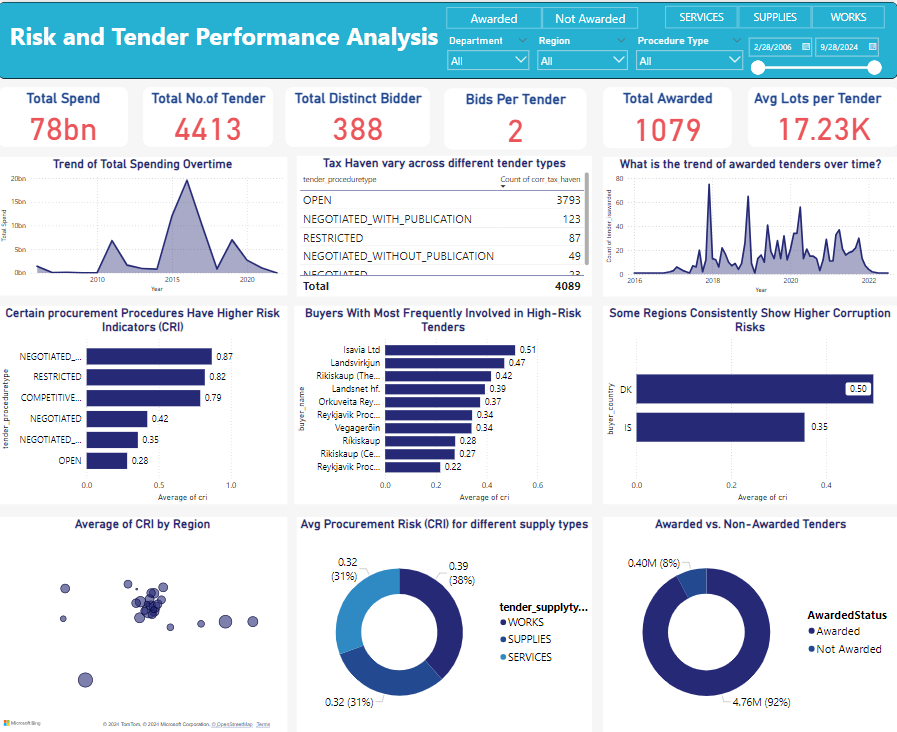
* The high award rate suggests an effective tendering process, where the majority of procurement activities proceed as planned.
* The smaller percentage of unawarded tenders could be due to factors such as inadequate bids, insufficient competition, non-compliance with requirements, or budgetary limitations.
* Addressing the reasons behind the unawarded tenders could improve the overall efficiency, ensuring that more projects are successfully completed and public funds are utilized effectively.

### 6.7 Dashboard

**Dashboard 1**



**Dashboard 2**



## 7. Public Procurement Analyst GPT

The Public Procurement Analyst GPT is an advanced AI tool specifically designed to automate and streamline the analysis of public procurement datasets. Using data from various global sources, such as the “Global Contract-level Public Procurement Dataset” published in *ScienceDirect*, this GPT efficiently processes procurement spending data from across the world. In this section, we explore how the tool operates and delivers insightful analyses in a user-friendly format.

**Initiating the Analysis**

Upon launching the GPT, users are given the option to either start a new analysis or continue with the existing dataset. By selecting "Start over," the tool prompts you to upload a new dataset or proceed with the current one. If a new dataset is uploaded, the system automatically checks for consistency, ensuring that the feature names and data types align with the original dataset format. Once validated, the GPT initiates its predefined analysis processes.

If the user chooses to proceed with the current dataset, the validation step is skipped, and the analysis begins immediately. The tool’s interactive design allows users to control the pace, guiding them through each stage of the analysis. By typing “yes,” users can move forward step-by-step, ensuring that the experience is both intuitive and responsive.

**Detailed Analysis and Visualization**

Each analysis conducted by the GPT is accompanied by professional-grade visualizations, offering a comprehensive look at key metrics and trends. These visual outputs are complemented by a concise report that details the findings, insights, and recommendations derived from the data. The combination of charts, graphs, and narrative explanations provides users with a clear understanding of the results, aiding in better decision-making.

Upon completing the first analysis, the system prompts the user with the option to continue to the next stage. By typing “yes,” the GPT seamlessly transitions to the next phase, keeping the user fully engaged in the process.

**Generating a Comprehensive Report**

One of the most valuable features of the Public Procurement Analyst GPT is its capability to produce a comprehensive report. By typing “create report,” the tool compiles the results of all analyses into a final document. This report includes detailed insights, strategic recommendations, and visual summaries, making it an indispensable resource for decision-makers and stakeholders involved in procurement.

The Public Procurement Analyst GPT is reshaping the way public procurement data is analyzed. By automating complex processes, it reduces manual effort, improves accuracy, and accelerates the overall decision-making workflow. For procurement professionals, this tool not only enhances efficiency but also promotes better governance and transparency in procurement practices.

**8. Conclusion**

The Global Public Procurement Dataset (GPPD) provides a comprehensive and invaluable resource for understanding and analyzing public procurement processes worldwide. Covering over 72 million contracts from 42 countries between 2006 and 2021, the dataset offers standardized and detailed information on government spending, including key attributes such as tender types, procurement entities, and contract values. The GPPD fills a crucial gap in global procurement data, enabling researchers, policymakers, and auditors to monitor and evaluate public procurement performance with unprecedented scope and depth.

The GPPD, with its extensive coverage and detailed information, is a powerful tool for enhancing transparency, efficiency, and accountability in public procurement. The innovative use of GPT and prompt engineering further maximizes its value, making complex data accessible and actionable. By providing robust and relevant findings, the GPPD supports efforts to improve governance, reduce corruption, and ensure the effective use of public funds. This comprehensive approach ultimately contributes to better governance and increased public trust in government spending practices.

## 9. Statistics

* **Total Contract Value:** Over USD 16.8 trillion, which is about 1.1% of global GDP annually.
* **Number of Contracts:** Over 72 million contracts involving around 1.8 million buyers and more than 10 million suppliers.
* **Corruption Risk Index:** A composite corruption risk index for each contract, enabling an objective assessment of corruption risks and facilitating comparisons across time, organizations, or countries.
* **Average Contract Value:** Around USD 230 million.
* **Number of Bids:** Over 1.2 billion bids submitted for the contracts.
* **Award Rate:** Around 70% of contracts were awarded.
* **Duration of Contracts:** The average duration of contracts is around 2.5 years.

**References**

Bosio, E., Djankov, S., Glaeser, E., & Shleifer, A. (2020). Public procurement in law and practice. NBER Working Paper No. 27188. Retrieved from https://www.nber.org/papers/w27188

Dimitri, N., Piga, G., & Spagnolo, G. (Eds.). (2006). Handbook of procurement. Cambridge University Press. Retrieved from https://www.cambridge.org/core/books/handbook-of-procurement/AC38AB1779D5DFA083D09C0EB4E12E1A

Dunning, J. H. (2000). The eclectic paradigm as an envelope for economic and business theories of MNE activity. International Business Review, 9(2), 163-190. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0969593199000359>

European Commission. (2018). Public procurement. Retrieved from <https://ec.europa.eu/growth/single-market/public-procurement_en>

Fazekas, M., & Tóth, B. (2016). The extent and evolution of corruption risks in Hungarian public procurement. European Journal on Criminal Policy and Research, 22(3), 429-448. Retrieved from <https://link.springer.com/article/10.1007/s10610-016-9307-7>

Flynn, A., & Davis, P. (2017). Explaining SME participation and success in public procurement using a capability-based model of tendering. Journal of Public Procurement, 17(3), 337-372. Retrieved from https://www.emerald.com/insight/content/doi/10.1108/JOPP-17-03-2017-B003/full/html

OECD. (2016). Government at a glance 2016. OECD Publishing. Retrieved from https://www.oecd.org/gov/government-at-a-glance-2016-9789264260016-en.htm

Piga, G., & Thai, K. V. (Eds.). (2007). The economics of public procurement. Palgrave Macmillan. Retrieved from <https://link.springer.com/book/10.1057/9780230307520>

Sánchez Graells, A. (2015). Public procurement and the EU competition rules. Bloomsbury Publishing. Retrieved from https://www.bloomsburyprofessional.com/uk/public-procurement-and-the-eu-competition-rules-9781849467340/

Thai, K. V. (2001). Public procurement re-examined. Journal of Public Procurement, 1(1), 9-50. Retrieved from <https://www.researchgate.net/publication/237423245_Public_Procurement_Re-examined>

Wittig, W. A. (2003). Public procurement and the development agenda. International Trade Centre. Retrieved from https://www.intracen.org/uploadedFiles/intracenorg/Content/Publications/Public%20Procurement%20and%20the%20Development%20Agenda.pdf

World Bank. (2020). Procurement in World Bank investment project financing: Goods, works, non-consulting and consulting services. World Bank Group. Retrieved from <https://openknowledge.worldbank.org/handle/10986/33354>

Arrowsmith, S., & Kunzlik, P. (Eds.). (2009). Social and environmental policies in EC procurement law: New directives and new directions. Cambridge University Press. Retrieved from https://www.cambridge.org/core/books/social-and-environmental-policies-in-ec-procurement-law/CBA56756C71D24CF4F3160B5EC3B0B0C

**Appendices**

**Appendix A: Data Fields Description**

1. **persistent\_id**: A unique ID for each tender, created by hashing the URL of the first publication about the tender.
2. **tender\_id**: An internal ID assigned to each tender during data processing.
3. **tender\_title**: The title of the tender.
4. **tender\_proceduretype**: The type of procedure used for the tender, standardized to common categories like Open, Restricted, Negotiated, etc.
5. **tender\_nationalproceduretype**: The procedure type as originally published, including jurisdiction-specific terms.
6. **tender\_isawarded**: Indicates if the tender was awarded (true or false).
7. **tender\_supplytype**: The type of purchase: supplies, services, or public works.
8. **tender\_biddeadline**: The final deadline for bid submissions.
9. **tender\_isjointprocurement**: Indicates if the tender is a joint procurement involving multiple public bodies.
10. **tender\_lotscount**: Number of lots (subsections) in the tender.
11. **tender\_recordedbidscount**: Number of unique bids recorded for the tender.
12. **tender\_isframeworkagreement**: Indicates if the tender is a framework agreement (a long-term agreement for multiple purchases).
13. **tender\_isdps**: Indicates if the tender is a Dynamic Purchasing System, a flexible way to manage procurement.
14. **tender\_contractsignaturedate**: The date the contract was signed.
15. **tender\_cpvs**: List of product codes for the items purchased, based on Common Procurement Vocabulary (CPV) codes.
16. **tender\_maincpv**: Main product code for the tender, based on CPV codes.
17. **tender\_iseufunded**: Indicates if the tender is funded by the European Union.
18. **tender\_selectionmethod**: The criteria for selecting the winning supplier: lowest price or most economically advantageous tender (MEAT).
19. **tender\_awardcriteria\_count**: Number of criteria used to evaluate the bids.
20. **tender\_cancellationdate**: The date the tender was canceled.
21. **cancellation\_reason**: The reason for the tender or contract cancellation.
22. **tender\_awarddecisiondate**: The date the award decision was made.
23. **tender\_estimatedprice**: The estimated price of the tender.
24. **tender\_finalprice**: The final price of the tender.
25. **lot\_estimatedprice**: The estimated price for each lot within the tender.
26. **bid\_price**: The price of a specific bid.
27. **tender\_corrections\_count**: Number of corrections made to the tender.
28. **lot\_row\_nr**: Unique identifier for each lot within a tender.
29. **lot\_title**: The title of the lot.
30. **lot\_status**: Indicates if the lot was awarded.
31. **lot\_bidscount**: Total number of bids submitted for the lot.
32. **lot\_validbidscount**: Total number of valid bids for the lot.
33. **lot\_electronicbidscount**: Number of bids submitted electronically for the lot.
34. **lot\_smebidscount**: Number of bids submitted by Small and Medium-sized Enterprises (SMEs) for the lot.
35. **lot\_updateddurationdays**: Latest duration of the lot or contract, in days.
36. **buyer\_id**: Main identifier of the buyer from the source documents.
37. **buyer\_masterid**: Unique identifier of the buyer, assigned during data processing.
38. **buyer\_name**: Name of the buyer.
39. **buyer\_nuts**: Regional code of the buyer, based on Nomenclature of Territorial Units for Statistics (NUTS).
40. **buyer\_city**: City of the buyer.
41. **buyer\_country**: Country of the buyer.
42. **buyer\_mainactivities**: Main activities of the buyer (e.g., health, education, public services).
43. **buyer\_buyertype**: Type of buyer (e.g., national authority, regional authority, public body).
44. **buyer\_postcode**: Postcode of the buyer.
45. **buyer\_nuts\_1**: First-level NUTS code of the buyer's location.
46. **buyer\_nuts\_2**: Second-level NUTS code of the buyer's location.
47. **buyer\_nuts\_3**: Third-level NUTS code of the buyer's location.
48. **buyer\_street**: Street address of the buyer.
49. **buyer\_url**: Website of the buyer.
50. **buyer\_email**: Email address of the buyer.
51. **buyer\_phone**: Phone number of the buyer.
52. **buyer\_contactName**: Contact person's name at the buyer's organization.
53. **buyer\_extra\_source\_id**: Other identifiers of the buyer from source documents.
54. **buyer\_sourceid\_type**: Type of other identifiers for the buyer.
55. **bidder\_id**: Main identifier of the bidder company from the source documents.
56. **bidder\_masterid**: Unique identifier of the bidder company, assigned during data processing.
57. **bidder\_name**: Name of the bidder company.
58. **bidder\_nuts**: Regional code of the bidder company, based on NUTS.
59. **bidder\_city**: City of the bidder company.
60. **bidder\_country**: Country of the bidder company.
61. **bidder\_postcode**: Postcode of the bidder company.
62. **bidder\_street**: Street address of the bidder company.
63. **bidder\_email**: Email address of the bidder company.
64. **bidder\_phone**: Phone number of the bidder company.
65. **bidder\_extra\_source\_id**: Other identifiers of the bidder company from source documents.
66. **bidder\_sourceid\_type**: Type of other identifiers for the bidder company.
67. **bidder\_url**: Website of the bidder company.
68. **bidder\_contactName**: Contact person's name at the bidder company.
69. **bidder\_nuts\_3**: Third-level NUTS code of the bidder company's location.
70. **bidder\_nuts\_2**: Second-level NUTS code of the bidder company's location.
71. **bidder\_nuts\_1**: First-level NUTS code of the bidder company's location.
72. **bid\_iswinning**: Indicates if the bid was a winning bid (true or false).
73. **bid\_issubcontracted**: Indicates if part of the contract will be subcontracted.
74. **bid\_subcontractedproportion**: The portion of the contract expected to be subcontracted.
75. **bid\_isconsortium**: Indicates if the bid is submitted by a consortium.
76. **source**: Source of the tender data.
77. **tender\_publications\_lastcontractawardurl**: URL of the last contract award announcement.
78. **tender\_publications\_firstdcontractawarddate**: Date of the first contract award announcement.
79. **notice\_url**: URL of the last call for tenders or contract notice publication.
80. **tender\_publications\_firstcallfortenderdate**: Date of the first call for tender announcement.
81. **tender\_year**: Year of the tender.
82. **tender\_addressofimplementation\_nuts**: Regional code of the tender implementation location, based on NUTS.
83. **tender\_description\_length**: Length of the tender description in characters.
84. **lot\_description\_length**: Length of the lot description in characters.
85. **tender\_personalrequirements\_length**: Length of the personal requirements for participation in characters.
86. **tender\_technicalrequirements\_length**: Length of the technical requirements for participation in characters.
87. **tender\_economicrequirements\_length**: Length of the economic requirements for participation in characters.
88. **currency**: Currency of the prices.
89. **tender\_digiwhist\_price**: Estimated final price of the tender, using various methods if the actual final price is not available.
90. **bid\_digiwhist\_price**: Estimated contract price, using bid price or lot estimated price if the bid price is missing.
91. **lot\_id**: Unique identifier for a lot, assigned during data processing.
92. **bid\_id**: Unique identifier for a bid, assigned during data processing.
93. **bid\_priceUsd**: Bid price converted to International USD.
94. **lot\_estimatedpriceUsd**: Estimated lot price converted to International USD.
95. **tender\_estimatedpriceUsd**: Estimated tender price converted to International USD.
96. **tender\_finalpriceUsd**: Final tender price converted to International USD.

**Filters**

1. **filter\_framework**: Filters out framework agreements unless the resulting contracts of minitenders are included.
2. **filter\_buyer/filter\_bidder**: Filters out rows with missing or erroneous buyer or bidder names.
3. **filter\_cancelled**: Filters out tenders that were canceled.
4. **filter\_opentender**: Deduplicates tenders from overlapping data sources, keeping only one instance based on certain criteria like value thresholds.
5. **filter\_year**: Filters for years where data quality is good and consistent.
6. **filter\_losingbids**: Filters out rows that refer to losing bids.
7. **filter\_ok**: A combination of filters to select non-duplicated awarded tenders/lots for analysis.

**Corruption Risk Indicators**

1. **corr\_singleb**: Indicates if a lot received only one bid (1 if true, 0 otherwise).
2. **corr\_proc**: Indicates if the tender procedure type is non-open (1 if true, 0 otherwise).
3. **submission\_period**: Number of days between the first call for tenders publication and the bidding deadline.
4. **corr\_subm**: Indicates if the submission period is unusually short (1 if true, 0 otherwise).
5. **corr\_nocft**: Indicates if there is no call for tenders publication (1 if true, 0 otherwise).
6. **decision\_period**: Number of days between the bidding deadline and the award decision.
7. **corr\_decp**: Indicates if the decision period is unusually short (1 if true, 0 otherwise).
8. **corr\_tax\_haven**: Indicates if the supplier is from a high financial risk country (1 if true, 0 otherwise).
9. **corr\_spending\_concentration**: Share of the total amount won by a specific supplier from a given buyer.
10. **cri (Composite Risk score)**: Average of the above risk scores, indicating overall corruption risk.