Public Tender Analysis and Visualization for Nova Scotia

CSCI6612 - Visual Analytics, Fall 2024

Md Samshad Rahman
B00968344
samshad@dal.ca

Table of Contents

Abstract

This project, Public Tender Analysis and Visualization for Nova Scotia, provides an easy-to-use platform to explore government tenders. It helps users analyze awarded tenders, track vendor payments, and investigate specific entities for opportunities or trends. The data comes from the Nova Scotia Open Data Portal and includes details like vendors, entities (organizations), tender types, amounts, and timelines. Each instance is one tender data. The project turns raw data into clear visualizations, making it easier for businesses, researchers, and policymakers to understand public procurement. By promoting transparency and fair competition, it encourages trust in public spending and supports research on economic trends.

Solving the Problem and Its Importance

Public procurement is a key part of good governance, involving the use of significant public funds. However, it often faces challenges like inefficiency, lack of transparency, and risks of corruption, which can erode public trust and hinder fair competition. In Nova Scotia, while tender data is publicly available, it is not presented in a way that is easy to explore or understand, making it harder for stakeholders to monitor spending or gain insights.

This project tackles these issues by turning complex tender data into simple, visual insights. The platform helps users:

1. Investigate awarded tenders to identify anomalies, inefficiencies, or potential corruption.
2. Gain insights into tendering practices, such as monetary amounts awarded and vendor participation.
3. Research vendors or entities for historical performance and future opportunities.

Importance of the Project

Transparent and efficient public procurement matters for several key reasons:

1. **Building Trust:** Clear and accessible information helps the public trust how government funds are being used.
2. **Encouraging Fair Competition:** Making tender practices more visible allows businesses to compete equally, driving innovation and economic growth.
3. **Promoting Accountability:** Highlighting spending trends and procurement patterns helps hold government entities accountable for their decisions.
4. **Supporting Better Decisions:** Organized data makes it easier for researchers, policymakers, and businesses to analyze trends, plan effectively, and develop strategies.

By turning raw data into actionable insights, this project ensures that public procurement in Nova Scotia is transparent, fair, and efficient. It helps create a system that everyone can trust.

Proposed Solution

Public procurement often lacks transparency, making it hard to track spending, spot inefficiencies, and ensure fair competition. This project addresses these issues by building a platform that uses visualization and machine learning to analyze awarded tenders in Nova Scotia. The platform helps users explore tender practices, uncover patterns, and gain meaningful insights to promote accountability and fairness.

Visualization View

The visualization view provides an interactive way to analyze tender data using four key tabs:

1. **Cluster Analysis:** Groups entities based on tender behavior, enabling comparisons between similar entities.
2. **Entity Analysis:** Focuses on individual entities to provide detailed insights into their tendering patterns.

Key Features:

- **Dynamic Filtering:** Explore data by clusters, entities, and tender categories (Goods, Services, Construction).
- **Modal Popups:** View detailed tender information by interacting with visual elements.
- **BERTopic Visualizations:** Includes word clouds, topic distribution plots, and interactive relationship plots to uncover trends and themes.
- **Bar and Line Charts:** Highlight trends in awarded amounts and vendor participation over time.

Machine Learning Module

The machine learning module analyzes tender data to uncover patterns and themes:

Clustering Analysis:

- Methods Used: K-Means and Agglomerative Clustering on entity names.
- Optimization: The Elbow Method determined 15 optimal clusters, with an additional manually curated "Health" cluster.
- Insights: Groups entities with similar tendering behaviors for targeted analysis.

Topic Modeling with BERTopic:

- Extracts key themes from tender descriptions.
- Visualizes topic distribution over time and categories, revealing shifts in procurement priorities.

Novel Features

Visualization Innovations

- **Cluster-Specific Insights:** Tailored visualizations highlight trends within grouped entities, revealing nuanced patterns.
- **Interactive Topic Exploration:** BERTopic-based plots let users dynamically explore thematic connections and topic trends.
- **Entity-Level Focus:** Filters allow precise analysis of individual entities’ tender behaviors and vendor relationships.

Machine Learning Innovations

- **Hybrid Clustering Approach:** Combines algorithms like K-Means and Agglomerative Clustering with domain expertise for meaningful clusters.
- **Dynamic Topic Insights:** BERTopic identifies evolving procurement priorities, offering strategic foresight into government spending patterns.

Impact and Importance

By merging advanced visualizations with machine learning, this project:

- **Enhances Transparency:** Makes tender data accessible and easy to understand, building trust in public procurement.
- **Supports Accountability:** Highlights inefficiencies and irregularities for better oversight and informed decisions.
- **Drives Research and Innovation:** Provides a rich dataset for analyzing economic trends and improving procurement strategies.

Ultimately, this project transforms raw data into actionable insights, ensuring Nova Scotia's tendering processes are fair, efficient, and transparent.

Justification and Explanation of the Machine Learning Approach

The machine learning approach focuses on Clustering Analysis and Topic Modeling with BERTopic, chosen to uncover patterns and insights in tender data. These methods align with the project's goals of enhancing transparency, accountability, and informed decision-making.

Clustering Analysis

Clustering groups entities based on their domain, helping stakeholders identify similarities and differences to better understand procurement practices.

Techniques Used:

1. K-Means Clustering
   - Chosen for its efficiency and ability to handle large datasets.
   - The Elbow Method determined 14 optimal clusters, ensuring meaningful group distinctions.
2. Agglomerative Clustering
   - Used to cross-validate K-Means clusters for robustness.
   - Provides hierarchical insights, showing relationships between smaller and larger clusters.

Why Clustering?

Clustering provides valuable insights by grouping entities with similar tendering behaviors. It helps in the following ways:

- **Comparative Analysis:** Enables stakeholders to compare similar entities, uncovering best practices or potential anomalies.
- **Prioritization:** Focuses attention on specific clusters, such as those with high spending or irregular patterns, for deeper analysis.
- **Customization:** Allows users to drill down into clusters or tender categories (e.g., Goods, Services, Construction) for tailored insights.

Topic Modeling with BERTopic

Purpose:

BERTopic extracts thematic insights from tender descriptions to reveal procurement priorities and focus areas.

Technique Used:

BERTopic

- Combines BERT embeddings (context-aware language model) with topic modeling to generate meaningful, coherent topics.
- Provides dynamic visualizations such as word clouds, topic distribution plots, and interactive topic relationship graphs.

Why BERTopic?

1. **Context-Aware Insights:** Captures nuanced relationships between words, making it ideal for analyzing domain-specific text like tender descriptions, unlike traditional methods (e.g., LDA).
2. **Dynamic Analysis:**
   - Tracks topic prevalence over time, uncovering evolving procurement priorities.
   - Categorizes topics by clusters for tailored insights into specific entities or vendors.
3. **Interactivity:** Allows users to visually explore topics, enhancing engagement and understanding

Justification for the Approach

Problem-Driven Selection:

- Clustering categorizes entities based on tender behavior.
- Topic modeling reveals the thematic focus of procurement activities.

Scalability:

- Both K-Means and BERTopic effectively handle large datasets, ensuring scalability for growing tender data.

Interpretability:

- Clusters provide clear groupings for stakeholders.
- BERTopic visualizations simplify complex topics for diverse audiences.

Impact:

- The combination of clustering and topic modeling bridges quantitative and qualitative analysis for a holistic understanding of tender data.

How the Approach Solves the Problem

- **Transparency:** Clustering reveals procurement patterns, while BERTopic highlights thematic trends, making processes easier to understand.
- **Accountability:** Helps identify inefficiencies or irregularities through entity groupings and thematic analysis.
- **Decision-Making:** Provides actionable insights for refining procurement strategies, allocating resources, and fostering fair competition.

By combining clustering and topic modeling, this approach offers a comprehensive analysis of Nova Scotia's tender data, aligning with the project's goal to promote an accountable and transparent procurement system.

Data Cleaning and Preprocessing

To ensure the dataset's reliability and usability, several cleaning and preprocessing steps were performed:

1. **Standardization of Vendor and Entity Names**:

   - Variations and inconsistencies in naming (e.g., "ABC Ltd." vs. "A.B.C Limited") were addressed through manual mapping to consolidate similar entries.
   - This reduced the complexity of the dataset significantly:
     a. Unique vendors reduced from ~12,594 to 5,600.
     b. Unique entities reduced from 225 to 215.

2. **Data Cleaning**

   Excluded records with insufficient or ambiguous information, such as:

   - Tenders with vague vendor names like "Multiple Vendors."
   - Missing descriptions or critical fields (e.g., awarded amounts).
   - Tenders with awarded amounts below $1,000, deemed insignificant for analysis.

3. **Feature Engineering**:

   - Created a DURATION column to calculate the number of days between TENDER_START_DATE and TENDER_CLOSE_DATE, offering insights into tendering efficiency.
   - Converted binary columns (e.g., GOODS, SERVICES) into numerical format (Y = 1, N = 0) to ensure compatibility with data analysis and machine learning algorithms.

These steps ensured the dataset was accurate, well-structured, and ready for in-depth analysis and visualization.

Justification and Explanation of Each Visualization View

This project's interface integrates two visualization views: Cluster Analysis and Entity Analysis. Both views are designed to provide insightful, actionable, and interactive data visualizations tailored to user needs.

1. Cluster Base View

i. Why was this view chosen?

- **Purpose:**
  - This view helps users explore tendering patterns across clusters of entities grouped by similar tendering behavior. It enables comparative analysis of procurement practices and spending trends within these clusters.

- **Reasoning:**
  - Clusters offer logical groupings of entities with similar procurement practices, making it easier to identify patterns and irregularities.
  - Users can gain insights into high-spending clusters, category-specific procurement, and vendor concentration.

ii. What does this view present to the user?

This view provides:

- **Cluster Selection and Category Analysis:**
  - Enables users to focus on clusters of interest and analyze procurement in specific categories (Goods, Services, Construction).

- **Tendering Dynamics Over Time:**
  - **Tenders by Year and Awarded Amount:** Highlights spending patterns, vendor participation, and awarded amounts over time.
  - **Cluster-Year Cumulative Awarded Amount:** Shows how spending evolves across clusters over time.
  - **Cluster-Year Average Awarded Amount:** Detects anomalies or trends in spending efficiency.

- **Vendor Insights:**
  - **Top Vendors by Tender Frequency:** Identifies vendors dominating procurement within specific clusters.
  - **Top Vendors by Awarded Amount:** Highlights vendors receiving the highest monetary awards, uncovering procurement dependencies.

- **Topic Modeling Visualizations:**
  - **Word Cloud:** Visualizes prevalent tender topics in clusters for quick thematic understanding.
  - **Topic Distribution and Visualization Plots:** Tracks shifts in procurement priorities or trends across clusters.

2. Entity Base View

i. Why was this view chosen?

- **Purpose:**
  - This view focuses on individual entities, allowing users to investigate specific government bodies' procurement practices.

- **Reasoning:**
  - Provides a detailed view of an entity's tendering behavior to assess accountability and transparency.
  - Supports vendor research, opportunity identification, and policy evaluation.

ii. What does this view present to the user?

This view highlights:

- **Entity-Specific Tendering Behavior:**
  - Allows users to explore an entity's procurement trends, including types of tenders issued and monetary spending patterns.

- **Cluster Mapping:**
  - Contextualizes the selected entity within its broader cluster, enabling comparative analysis.

- **Vendor-Focused Insights:**
  - **Top Vendors by Tender Frequency:** Identifies vendors frequently awarded tenders by the entity.
  - **Top Vendors by Awarded Amount:** Highlights high-value vendors to detect procurement dependencies or favoritism.

- **Temporal Trends in Procurement:**
  - **Tenders by Year and Awarded Amount:** Visualizes an entity's spending trends over time.
  - **Topic Distribution Over Time:** Tracks how procurement priorities evolve, aiding in the detection of shifts in policy or focus.

- **Topic Modeling Visualizations:**
  - **Word Cloud:** Displays dominant tender topics for the entity.
  - **Topic Distribution Over Time:** Offers a longitudinal view of procurement themes.

Why These View Work Together

- The Cluster Analysis View provides a macro perspective, helping users understand trends and patterns across groups of entities.
- The Entity Analysis View offers a micro perspective, delivering granular insights into individual entities' behaviors.

Together, these views balance breadth and depth, empowering users with comprehensive and actionable insights into Nova Scotia's tendering practices.

Tools, Libraries, and External Resources Used

The project relied on a combination of advanced tools and libraries to perform data analysis, clustering, and visualization, as well as to integrate interactive elements for effective user engagement.

1. Programming Language: Python

- **Purpose:**
  Python served as the core programming language for all aspects of the project, including data manipulation, analysis, machine learning, and visualization. Python's vast ecosystem of libraries allowed seamless integration of various tasks into one cohesive pipeline.

2. Libraries and Tools:

Dash and Plotly

- **Purpose:**
  Dash, built on top of Plotly, was used to create the interactive web-based dashboards. Plotly provided the tools for rendering various chart types (e.g., bar charts, line charts, word clouds, etc.), while Dash allowed for the integration of interactivity and real-time user input.

- **Changes/Additions:**
  Customization was applied to Dash components, including the use of callback functions to link user inputs (e.g., filtering options, cluster selections) to dynamic chart updates.
  - *Example:* Added dynamic filters such as cluster selection, vendor-year filtering, and category-specific analyses.
  - Integrated modal popups for detailed tender views when users interact with visual elements.

Pandas

- **Purpose:**
  Pandas was used for data preprocessing and manipulation, including data cleaning, handling missing values, and reshaping the data to fit the needs of visualization and analysis.

- **Changes/Additions:**
  - Preprocessed the tender data to group by clusters and entities.
  - Created new variables such as "award duration" by calculating differences between the TENDER_START_DATE and AWARDED_DATE.
  - Implemented custom aggregation functions to calculate total and average amounts by year, vendor, and cluster.

Scikit-learn

- **Purpose:**
  Scikit-learn was used for implementing clustering algorithms (K-Means and Agglomerative Clustering) and other preprocessing techniques. It allowed for the grouping of entities based on similar tendering behavior, as well as the extraction of optimal cluster counts.

- **Changes/Additions:**
  - The **Elbow Method** was used to determine the optimal number of clusters.
  - Integrated both **K-Means** and **Agglomerative Clustering** algorithms to classify entities into meaningful groups based on their tendering characteristics.
  - Used scaling and encoding techniques to preprocess the data (e.g., one-hot encoding for categorical features).

BERTopic

- **Purpose:**
  BERTopic was employed for advanced topic modeling of tender descriptions, enabling the extraction of thematic patterns from text data. This allowed the visualization of topics and their distribution within different clusters and categories of tenders.

- **Changes/Additions:**
  - Implemented **BERTopic** for topic extraction from the tender descriptions, capturing the most significant themes associated with different tenders.
  - Integrated interactive visualizations like **word clouds** and **topic distribution plots** to represent the extracted topics.
  - Used BERTopic's **visualization tools** to create dynamic plots that track topic evolution over time and across different clusters or entities.

3. External Resources:

Datasets

- **Tender Data**: The primary data used in the project came from Nova Scotia's publicly available tender datasets.
  - This data was pre-processed using Python to remove any irrelevant or missing information, ensuring consistency and clarity for analysis.

By combining these tools, libraries, and techniques, the project created an intuitive, interactive interface for analyzing Nova Scotia's awarded tenders, allowing users to delve into patterns, trends, and relationships within the data.

Challenges and Limitations

The project encountered several challenges related to data quality, clustering, and scope:

- **Data Quality Issues**:
    - **Inconsistent Vendor and Entity Names**:
        - The dataset contained numerous typographical errors, variations, and inconsistencies in vendor and entity names. For instance, a single vendor might appear under multiple names due to formatting differences or abbreviations.
        - Resolving these discrepancies required significant manual effort, including mapping approximately 12,594 vendor entries into 5,600 standardized names and 225 entity entries into 215.

    - **Missing or Ambiguous Data**:
        - Some records had vague vendor descriptions (e.g., “Multiple Vendors”) or were missing critical fields like tender descriptions and awarded amounts.
        - These records were removed to maintain the integrity of the analysis, though it potentially limited the dataset’s comprehensiveness.

Future Work

While the project is effective in its current form, several enhancements and expansions could significantly improve its functionality, user-friendliness, and impact:

1. Scheduler-Driven Hosting:

- **Automatic Data Updates**:
  Tender data is regularly updated, and integrating a scheduler-driven hosting system would enable automatic fetching of the latest data each month. This ensures the dashboard remains up-to-date without manual intervention.

2. Enhanced Interactivity:

- **Entity and Vendor Management**:
  Introduce features for users to directly add or update entity and vendor details through the dashboard. This would maintain the system's relevance as new data becomes available or as entities and vendors evolve over time.

- **Dynamic Reassignments**:
  Allow users to reassign entities to different clusters dynamically. This functionality would enhance cluster-based analysis accuracy, especially when new insights or data emerge.

3. AI-Driven Insights:

- **Automated Tender Summaries**:
  Integrate Large Language Models (LLMs) to generate concise summaries of tender topics based on BERTopic outputs. This feature would provide users with quick, actionable insights without requiring deep exploration of the data.

- **Interactive Q&A**:

Incorporate natural language processing (NLP) features to allow users to ask specific questions, such as:

  - "Which vendors received the highest awards in the last five years?"
  - "What are the trends in construction tenders?" These insights could be generated in real-time, enhancing the user experience and making the platform more interactive.

- **Predictive Analytics:**
  Develop machine learning models to predict future tender amounts or durations based on historical data. This would help organizations identify procurement trends and prepare for upcoming procurement needs.

This project lays the groundwork for a robust procurement analysis tool. By implementing these proposed enhancements, the platform could become a vital resource for stakeholders involved in public procurement, supporting transparency, accountability, and strategic decision-making.

Conclusion

The "Public Tender Analysis and Visualization for Nova Scotia" project successfully transforms complex public tender data into actionable insights through a transparent and interactive platform. By leveraging advanced visualization and clustering techniques, the platform empowers various stakeholders, including citizens, researchers, and businesses, to:

- **Investigate Entities**: Explore procurement practices, evaluate transparency, and assess public spending across different entities.
- **Analyze Vendor Performance**: Gain detailed insights into vendor activity and awarded amounts, identifying performance trends and contributions.
- **Understand Trends**: Visualize tendering patterns over time to detect inefficiencies, irregularities, or opportunities for improvement in procurement practices.

The project underscores the critical role of transparency and accountability in public procurement. Its intuitive interface ensures accessibility for both technical and non-technical users, enabling them to extract meaningful insights from the data. By employing techniques such as BERTopic for topic modeling and clustering for grouping entities, the project demonstrates how data-driven approaches can enhance decision-making and improve procurement efficiency.

This initiative highlights the potential of modern data analysis tools to foster a fairer, more efficient, and transparent procurement system in Nova Scotia.

References

[1] Government of Nova Scotia, "*Awarded Public Tenders*," [Online]. Available: https://data.novascotia.ca/Procurement-and-Contracts/Awarded-Public-Tenders/m6ps-8j6u/about_data. [Accessed: Dec. 11, 2024].

[2] Plotly Technologies Inc., "*Plotly,*" [Online]. Available: https://plotly.com/. [Accessed: Dec. 11, 2024].

[3] Plotly Technologies Inc., "*Dash,*" [Online]. Available: https://dash.plotly.com/. [Accessed: Dec. 11, 2024].

[4] Scikit-learn Developers, "*Scikit-learn: Machine Learning in Python,*" [Online]. Available: https://scikit-learn.org/stable/. [Accessed: Dec. 11, 2024].

[5] The Pandas Development Team, "*Python Pandas,*" [Online]. Available: https://pandas.pydata.org/. [Accessed: Dec. 11, 2024].

[6] Python Software Foundation, "*Python Programming Language,*" [Online]. Available: https://www.python.org/. [Accessed: Dec. 11, 2024].