**Exploring Government Accountability**

**Through Open Data and Data Mining**

Sean Kuehl, Mason Ward, Dominic Pham

School of Applied Computer Science & Information Technology, Conestoga College

PROG74030: Data Mining

Dr. Sara Ross-Howe

August 16th, 2024

**Table of Contents**

[​ Abstract 3](#__RefHeading___Toc737_2535903744)

[​ Datasets and Initial Preparation 3](#__RefHeading___Toc739_2535903744)

[​ Overview 3](#__RefHeading___Toc741_2535903744)

[​ Description of Travel Expenses Data 3](#__RefHeading___Toc743_2535903744)

[​ Cleaning of Travel Expenses Data 4](#__RefHeading___Toc745_2535903744)

[​ Description of Hospitality Expenses Data 5](#__RefHeading___Toc747_2535903744)

[​ Cleaning of Hospitality Expenses Data 6](#__RefHeading___Toc749_2535903744)

[​ Combination Into Data warehouse 6](#__RefHeading___Toc751_2535903744)

[​ Loading Data Warehouse Into A Relational Database 8](#__RefHeading___Toc753_2535903744)

[​ Data Cleaning 9](#__RefHeading___Toc755_2535903744)

[​ Data Preparation 9](#__RefHeading___Toc757_2535903744)

[​ Cluster Analysis 10](#__RefHeading___Toc759_2535903744)

[​ Overview 10](#__RefHeading___Toc834_1196853497)

[​ Correlation Analysis using Heatmap 10](#__RefHeading___Toc761_2535903744)

[Centroid-Based Clustering: K-Means 12](#__RefHeading___Toc763_2535903744)

[​ Density-Based Clustering: DBSCAN 14](#__RefHeading___Toc765_2535903744)

[​ Discussion of Clustering Approaches 15](#__RefHeading___Toc767_2535903744)

[​ Decision Tree Analysis 16](#__RefHeading___Toc769_2535903744)

[​ Overview 16](#__RefHeading___Toc836_1196853497)

[​ Analysis 16](#__RefHeading___Toc838_1196853497)

[​ Association Mining 17](#__RefHeading___Toc771_2535903744)

[​ Overview 17](#__RefHeading___Toc840_1196853497)

[​ Itemset Frequency 18](#__RefHeading___Toc773_2535903744)

[​ Association Rules 18](#__RefHeading___Toc775_2535903744)

[​ Classification Analysis 20](#__RefHeading___Toc777_2535903744)

[​ Overview 20](#__RefHeading___Toc888_3591311785)

[​ Analysis 20](#__RefHeading___Toc890_3591311785)

[​ Classification Results 22](#__RefHeading___Toc779_2535903744)

[​ REFERENCES 23](#__RefHeading___Toc781_2535903744)

## Abstract

We all pay taxes, but how many of us can really say where exactly our money is going? We could not find an easy answer to this question, so we set out to find an answer through this analysis. One of the ways our tax dollars are spent is on services like healthcare, emergency services and the like. While these are a major source of government expenditure, the exact amount spent on them is largely a matter of politics and generally speaking, money spent on these topics produces greater benefits than it costs. The source of government expenditure we will be examining is discretionary spending, where some money may be put aside for these expenses but the exact amount spent can vary greatly and is largely up to individuals who are not directly accountable to the people. We used Government of Canada Open Data on hospitality and travel expenses for government agencies which are examples of discretionary spending. We prepared and performed analysis on this data to try and find trends and patterns we could use to provide actionable steps to help improve efficiency and reduce waste and corruption in discretionary government expenditure. We used several different methods to analyze and pull insights from the data including clustering, association rules and classification methods among others. This analysis was used to inform a list of recommendations for how to improve the processes involved with discretionary spending to better serve Canadian taxpayers.

## Datasets and Initial Preparation

### Overview

The datasets used were the Proactive Disclosure of Travel Expenses[1] and Proactive Disclosure of Hospitality Expenses[2]. These datasets are found on the Government of Canada Open Data Portal[3]. Due to the nature of these datasets, they are updated daily, so it is worth mentioning that the versions used for analysis were retrieved on July 19th, 2024. This data belongs to the Government of Canada and is used under the Open Government License[4]. These datasets share several columns in common which we will use to combine the two datasets into a single data warehouse.

### Description of Travel Expenses Data

This dataset contains 116,543 entries and has 19 attributes. Each entry in this dataset contains information on a travel expense made by a member of the public service while performing their duties. It should be noted that the attributes are not represented in the table exactly as they are in the raw data for clarity reasons. Descriptions of these attributes are as follows(Table 1).

Table 1

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| Reference Number | The unique identifier for the expense entry |
| Disclosure Group | This is the group to which the person traveling belongs. Which is either “MPSES” for Minister/Ministerial adviser/Ministerial staff/Parliamentary Secretary/Exempt Staff or “SLE” for Senior officer or employee |
| Title(English) | The official title of the person travelling in English |
| Title(French) | The official title of the person travelling in French |
| Name | The name of the person traveling |
| Purpose of Travel(English) | The provided reason for travel in English |
| Purpose of Travel(French) | The provided reason for travel in French |
| Travel Start Date | The start date of the individual’s travel |
| Travel End Date | The end date of the individual’s travel |
| Place Visited(English) | The location visited in English |
| Place Visited(French) | The location visited in French |
| Airfare | The cost of airplane tickets and similar |
| Other Transportation | The cost of other transportation; like rental cars for example |
| Lodging | The cost of accommodations |
| Meals and Incidentals | The cost of meals and the like |
| Other Expenses | The cost of any additional expenses not belonging to one of the previous kinds |
| Total Amount | The total cost of all expenses |
| Additional Comments(English) | Any additional notes made about the expense in English |
| Additional Comments(French) | Any additional notes made about the expense in French |
| Owner Org | The Government organization to which the individual belongs |
| Owner Org Title | The official title of the Government organization to which the individual belongs |

### Cleaning of Travel Expenses Data

The *Airfare*, *Other Transportation*, *Lodging*, *Meals and Incidentals* and *Other Expenses* attributes were dropped from the dataset. These attributes were dropped because they have no corresponding attribute in the other dataset and they are represented in the *Total Amount* attribute which is present in the other dataset. The *Name* attribute had values in several different formats. A name could be written in the form “Sean Kuehl”, “Kuehl, Sean” or “Sean L. Kuehl” for example. There were also some capitalization inconsistencies across the names. These different formats could pose an entity-identification issue later, so these formats were reduced into only the “Sean Kuehl” format and all names were converted to lowercase. The *Place Visited* attributes also posed an issue and contained “Toronto”, “Toronto, ON”, “Toronto, ON, Canada”, “Toronto, Ontario, Canada”, “Toronto (Ontario) Canada” or even “Toronto, On, Canada” formats. There were too many format variations to allow for a single deterministic rule to reduce them all down to a single format. In this case, only a lowercase filtering was applied to solve the “Toronto, ON”, “Toronto, On” inconsistency and any other capitalization issues. Producing a solution that converted all variations down to a single format would have been far more complex and required too many resources to fit within the project timeline.

### Description of Hospitality Expenses Data

The hospitality expenses dataset has 59,023 entries and 18 attributes. Each entry in this dataset contains information on a hospitality(purchasing of beverages, etc.) expense made by a member of the public service while performing their duties. These attributes are not represented exactly as they are in the raw data in the table for clarity reasons. The attributes are as follows(Table 2).

Table 2

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| Reference Number | The unique identifier for the expense entry |
| Disclosure Group | This is the group to which the person traveling belongs. Which is either “MPSES” for Minister/Ministerial adviser/Ministerial staff/Parliamentary Secretary/Exempt Staff or “SLE” for Senior officer or employee |
| Title(English) | The official title of the individual making the expense in English |
| Title(French) | The official title of the individual making the expense in French |
| Name | The name of the individual making the expense |
| Purpose of Hospitality Activity(English) | The provided reason for the expense in English |
| Purpose of Hospitality Activity(French) | The provided reason for the expense in French |
| Start Date | The start date of the activity associated with the expense |
| End Date | The end date of the activity associated with the expense |
| Municipality Where the Activity Took Place(English) | The town or city where the activity associated with the expense took place in English |
| Municipality Where the Activity Took Place(French) | The town or city where the activity associated with the expense took place in French |
| Name of Commercial Establishment or Vendor Involved In the Hospitality Activity(English) | The store(s) or business(es) who provided services for the activity in English(i.e. the hotel where an event takes place) |
| Name of Commercial Establishment or Vendor Involved In the Hospitality Activity(French) | The store(s) or business(es) where purchases as part of the expense were made in French(i.e. the hotel where an event takes place) |
| Attendees(Government of Canada Officials) | How many Government of Canada officials were present at the activity |
| Attendees(Guests) | How many non-Government of Canada officials were present at the activity |
| Total Cost | The total cost of all expenses involved with this activity |
| Additional Comments(English) | Any additional comments provided in English |
| Additional Comments(French) | Any additional comments provided in French |
| Owner Org | The Government organization the individual making the expense belongs to |
| Owner Org Title | The official title of the Government organization the individual making the expense belongs to |

### Cleaning of Hospitality Expenses Data

The *Name of Commercial Establishment or Vendor Involved In the Hospitality* *Activity* attributes, *Attendees(Government of Canada officials)* and *Attendees(Guests)* attributes were dropped because they have no corresponding attribute in the other dataset. The *Name* attribute in this dataset has similar issues to the *Name* attribute in the Travel Expenses dataset. This attribute was cleaned in the same fashion(See [Previous](#_Cleaning_of_Travel)). The *Municipality Where the Activity Took Place* attributes have some issues that need cleaning. Often, there are store or business names mixed in with the places. For example, “Ottawa, ON (Marcello's)”, or sometimes multiple locations in the form “Ottawa, ON (Costco, Sobeys, Loblaw)” or even “Super C, Costco, IGA, Art-is-in, Rideau Bakery, Sconewitch, Ottawa, ON". These formatting issues are on top of the same location format issues present in the previous dataset’s location attribute. This extra data added to the location is too varied and inconsistent to warrant keeping. This being said, the issue is too complex to resolve down to a single format with a single deterministic rule. For this reason, only the “Ottawa, ON (Marcello's)” case will be addressed by removing the content in brackets. A lowercase conversion was applied to all data contained within the attribute to resolve any capitalization issues present.

### Combination Into Data warehouse

The cleaned versions of the datasets were combined into a single data warehouse. This single data warehouse has several attributes that are common across the parent datasets along with a new attribute, “Expense Type” to denote which parent table the expense came from since the identifying information is not maintained through the merger into a data warehouse. The attributes are not shown exactly as they are in the raw data for clarity purposes. The attributes contained within the data warehouse are as follows(Table 3)

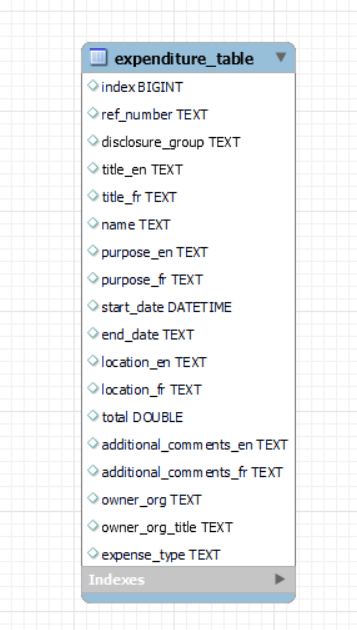
Table 3

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| Reference Number | The unique identifier for the expense entry |
| Disclosure Group | This is the group to which the person traveling belongs. Which is either “MPSES” for Minister/Ministerial adviser/Ministerial staff/Parliamentary Secretary/Exempt Staff or “SLE” for Senior officer or employee |
| Title(English) | The official title of the individual making the expense in English |
| Title(French) | The official title of the individual making the expense in French |
| Name | The name of the individual making the expense |
| Purpose(English) | The provided purpose for the expense in English |
| Purpose(French) | The provided Purpose for the expense in French |
| Start Date | The start date of the activity associated with the expense |
| End Date | The end date of the activity associated with the expense |
| Location(English) | The location where the activity associated with the expense took place in English |
| Location(French) | The location where the activity associated with the expense took place in French |
| Total | The total of all expenses involved with the activity |
| Additional Comments(English) | Any additional information provided about the expense in English |
| Additional Comments(French) | Any additional information provided about the expense in French |
| Owner Org | The Government organization the individual making the expense belongs to |
| Owner Org Title | The official title of the Government organization the individual making the expense belongs to |
| Expense Type | The type of expense and the parent table the expense belonged to. Either “Travel” or “Hospitality” |

## Loading Data Warehouse Into A Relational Database

In order to more efficiently store and access our data warehouse, we chose to store our data in a MySQL relational database. Tests were conducted to ensure the validity and accessibility of the data in the database. An Entity Relationship Diagram was also produced to assist with validation and analysis tasks(Figure 1).

Figure 1: Entity Relationship Diagram of the Data Warehouse



Note: Own Work

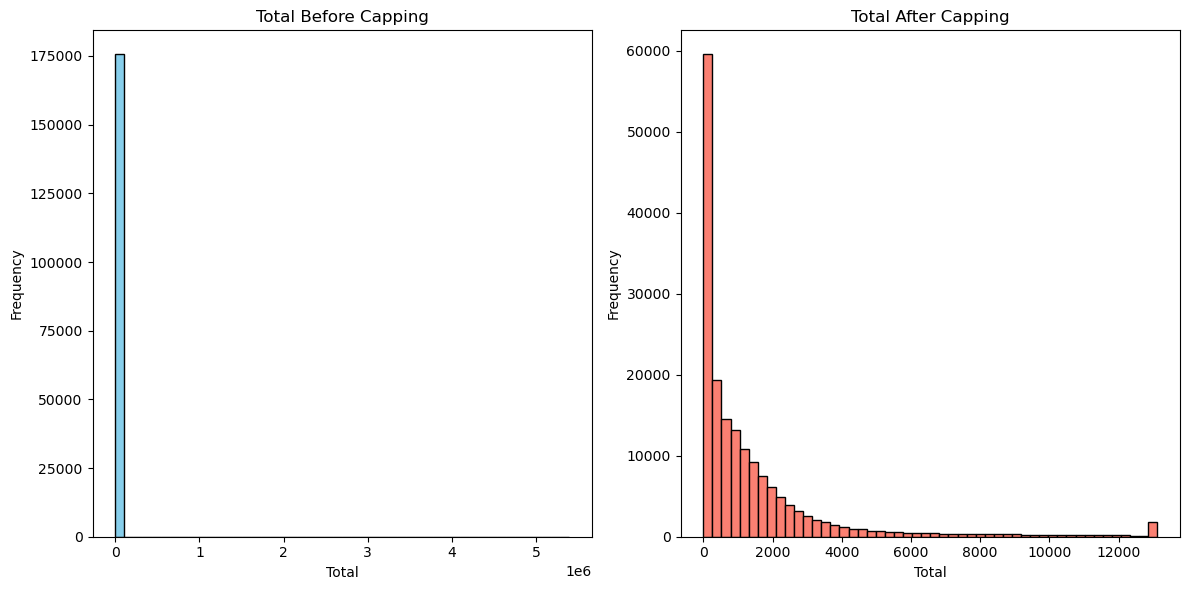
## Data Cleaning

Duplicates and missing values were addressed in the data warehouse as a part of data cleaning. First, duplicates were removed and a validation step was performed to ensure that the removal was successful. Once the duplication removal was complete, the missing values in the data were addressed. A significant amount of missing values were found across multiple columns, so we decided to use an imputation method as opposed to dropping the significant amount of data. For imputing into the numerical columns, the mean was chosen as it minimizes the sum of squared difference between the original and imputed values. For the categorical attributes except for *Name*, imputing based on the mode, or most common entry was chosen. Imputing with “Unknown” was chosen for the *Name* attribute as this provided more meaningful context for the analysis than imputing the entry with mode.

## Data Preparation

Outliers in the data were detected and mitigated and the data was encoded and binned appropriately to assist in future analysis. The Z-Score, IQR and Isolation Forest methods were used to detect outliers in the numerical attributes. All outliers detected were capped in order to give us the largest amount of data for analysis while still reducing the effect of outliers on that analysis. See the results below(Figure 2)

Figure 2: Effect of Outlier Capping On the Data



Note: Own Work

In addition, nominal attributes were integer encoded and were binned based on a “Low”, “Medium” and “High” ranges principal. For the date attributes, they were binned by extracting the year. The binning was performed by adding additional binned attributes to the data. Some date values were invalid and caused issues with binning. Since there were exceedingly few of these values (only 4), these were removed from the data.

## Cluster Analysis

### Overview

Our dataset is relevant and our aspirations noble, but can we actually use it to help improve government accountability? The first step to answering this question is performing cluster analysis. If we can find meaningful groups in our data, we may be able to explore them further and pull insights from them. These insights could lead us to finding odd expenses.

### Correlation Analysis using Heatmap

Using a heatmap, we conducted a feature correlation analysis to identify relevant features in the data and reduce features that are redundant. For this phase of the analysis, we conducted two iterations of the heatmap on the features of the dataset. The first iteration focuses on reducing feature redundancy and the second iteration focuses on selecting relevant features.

For the first iteration of our feature correlation heatmap, we identified that the English and French versions of certain columns such as purpose, location, etc. they both have equal value of correlation for that column, therefore dropping either the English or French version of those columns will not hinder our insights for finding potential patterns and this remove data redundancy. Another notable insight we gained from this heatmap is that for the start and end year columns, we have the non-binned and binned versions for these columns however we observe that the binning does not have any significant improvement that contribute to feature correlations when these columns are binned. Hence it seems appropriate to drop the binned versions of the columns. For the expense type column in the heatmap, it is showing up as blanks as the data itself had no relation to the other features in the dataset and it is decided that it should be removed as well as it would not provide any value. Lastly the reference number was identified as being a random numerical value used for referring to the entry and would provide no meaningful patterns or relations to the other features, hence it is removed as well.

1. Figure 3: Feature Correlation Heatmap (Initial version)

A screenshot of a computer code

Description automatically generated

Note: Own Work

For the second iteration of our feature correlation heatmap, we generated another heatmap with the redundant features that we identified in the first iteration removed. This provides a clear and concise heatmap in comparison to the first iteration of the heatmap where this version can help us identify complex relationships between the features in the dataset easier. For this version, we can now focus on relevant feature selection and observe how each feature relates to one another. We can identify from this heatmap that the columns *location*, *purpose*, *name*, and *title* have weak relationships with other features in the dataset and decided that these features would not provide any significant insight or patterns that would be useful for our clustering analysis in the next step.

1. Figure 4: Feature Correlation Heatmap (Revised version)

A colorful squares with black text

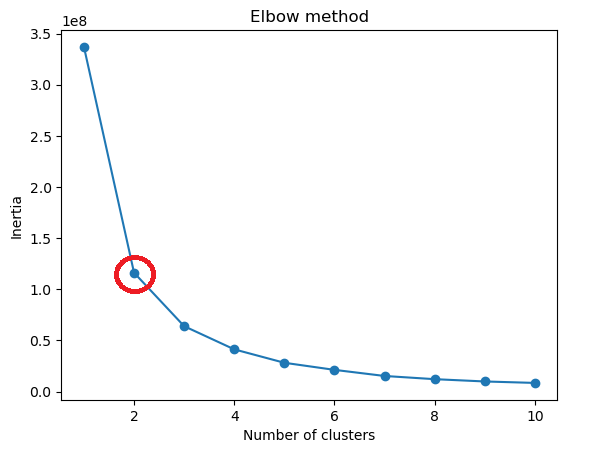
Description automatically generated with medium confidence

Note: Own Work

### Centroid-Based Clustering: K-Means

As part of our clustering analysis, we conducted a centroid-based clustering approach using K-Means as our clustering method. For this method, we started with determining the optimal value for k, the number of clusters that will be used. We identified that the optimal k value is 2.

1. Figure 5: Determining Optimal k value using Elbow Method



Note: Own Work

Using the k value that we identified using the elbow method, we conduct the K-Means clustering on the data in the *owner\_org* and *owner\_org\_title* columns as they are negatively correlated with one another and should be evenly dispersed on the plot.

1. Figure 6: K-Means Clustering

A chart with purple and yellow dots

Description automatically generated

Note: Own Work

### Density-Based Clustering: DBSCAN

1. For the next part of our clustering analysis, we conducted a density-based clustering approach using DBSCAN as our clustering method. For this method, the clustering is conducted on the same data as the previous clustering method.
2. Figure 7: DBSCAN Clustering

A chart with colored dots

Description automatically generated

Note: Own Work

### Discussion of Clustering Approaches

Based on observing the results of the two clustering approaches, K-Means clustering was efficient at clustering the data in the *owner\_org* and *owner\_org\_title* columns. It was able to clearly identify and create two clusters in the data that can be seen in the K-Means clustering plot. DBSCAN found more than four distinct clusters occupying different ranges from the top to the bottom of the graph. The clusters identified by DBSCAN are interesting and would provide notable insights for exploring patterns and finding trends. For these reasons, DBSCAN is the better clustering method for this dataset as it is more robust and capable of identifying patterns and trends than K-Means.

## Decision Tree Analysis

### Overview

Now that we know our expenses can be separated into several distinct groups, we need to find out what the meaning is behind them, if any. If there are reasons for these groupings then we can use these patterns to make inferences about our data and move to the next step in our process towards a tool for accountability.

### Analysis

Using the heatmap generated in the previous steps in our analysis, we can identify that the *owner\_org* column has strong correlation with other features in the dataset, hence we will use this column for our decision tree analysis where we will predict the values for this column. The trained a decision tree classifier on the data from the *owner\_org* column and achieve an accuracy of about 70%. The result of the full decision tree can be seen in the figure below.

Figure 8: Full Decision Tree Analysis (Depth=5)

A diagram of a computer network

Description automatically generated

Note: Own Work

To further examine the decision tree, we also generated a decision tree with a lower tree depth so that we can better analyze the decision that is made in the tree nodes. The figure below is an example of a reduced decision tree that provided more readability for us to better interpret the decision nodes in the tree as the full decision tree was difficult to read the decision nodes.

Figure 9: Reduced Decision Tree Analysis (Depth=1)

A diagram of a number

Description automatically generated

Note: Own Work

## Association Mining

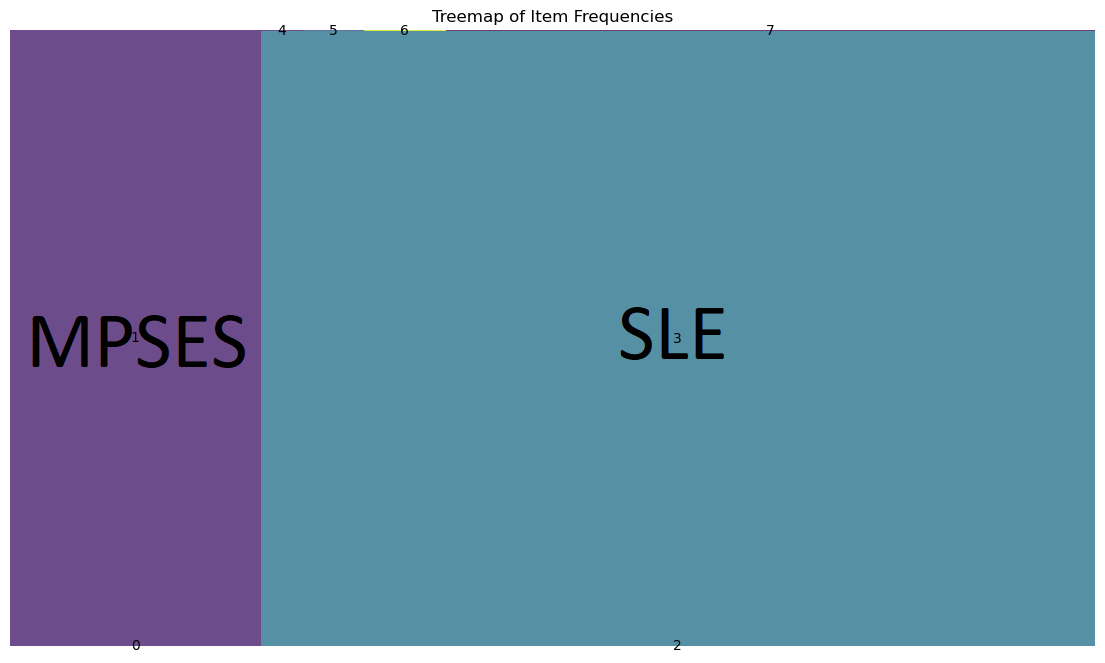
### Overview

We now know that there are meaningful patterns to our data. Let’s explore these further to gain a better understanding of them and possibly find more patterns. For this analysis we employed association mining to explore the relationship between disclosure groups and owner organization title. This can provide insight to the context of our analysis of ministerial travel and hospitality expense as this can help to identify patterns and association between disclosure groups in an organization and how the expenses are related to the disclosure groups.

### Itemset Frequency

For the frequency count of relevant itemset, we conducted frequency counts of disclosure groups among the organization titles. With the frequency count of the disclosure groups, we then visualize the frequency count with a tree map (Shown in Figure 10). From the tree map visualization, we can see that disclosure group 1 (MPSES) and 3 (SLE) are dominant across the organization. The disclosure groups were required to be encoded for our data processing hence they are represented as numerical in the data and the visualization. This indicates that a significant portion of travel and hospitality expenses are associated with senior officers and employees.

Figure 10: Tree map visualization of Disclosure Groups Frequency



Note: Own Work

### Association Rules

For the association rules, we generated rules set to specific thresholds for the lift, confidence, and support to identify interesting rules between the disclosure group and the organizations. We intended to further investigate what kind of organizations have certain disclosure groups associated with them.

Figure 11: Table of Interesting Rules

A screen shot of a computer screen

Description automatically generated

Note: Own Work

Some interesting rules we identified is that Owner Organization Title 52 (Global Affairs Canada) is highly associated with the disclosure group SLE, as indicated by the confidence metric of 83.55%. We can observe that transactions from senior officers and employees are likely to belong to Global Affairs Canada. Another notable rule is that Owner Organization Title 59 (Innovation, Science and Economic Development Canada) is associated with MPSES indicated by the confidence metric of 60.22%. We can observe that transactions of the disclosure group MPSES have a higher chance of the transaction belonging to Innovation, Science and Economic Development Canada. A few rules share a common confidence metric where the confidence is 100%, indicating that a specific disclosure group’s instances are associated with a specific organization. The following interesting rules have 100% confidence:

(Owner Organization Title: 19 (Canadian Heritage)) → (Disclosure Group: 3 (SLE))

(Owner Organization Title: 71 (National Research Council Canada)) → (Disclosure Group: 3 (SLE))

(Owner Organization Title: 108 (Royal Canadian Mounted Police)) → (Disclosure Group: 3 (SLE))

(Owner Organization Title: 83 (Office of the Prime Minister)) → (Disclosure Group: 1 (MPSES))

To further expand, these rules indicate that an organization consists of solely one disclosure group. This can be helpful with the analysis when required to focus on specific disclosure groups when looking at hospitality and travel expenses. We can narrow the search space of the organization if the organization does not include the disclosure group that we wish to focus on. The interesting rules overall add value to our analysis, as looking at the distribution of the disclosure groups among the organization can indicate patterns in how different discourse groups within the organization are linked to travel and hospitality expenses. The high confidence and lift values seen in the interesting rules for certain organizations can suggest that some disclosure groups for certain organizations require more travel and hospitality activities which can be flagged for further scrutiny for necessity and efficiency.

## Classification Analysis

### Overview

Now that we have well established meaningful patterns and explored our dataset in-depth, it’s now time to use this work to help improve the accountability of government discretionary expenditure. We have decided to create a classification model based on the information gained through our analysis. This model will be used to predict high-cost expenses so that we can predict, scrutinize, audit and track these expenses that may have room for efficiencies and improvements. We envision this tool could aid auditors on selecting government expenses to audit.

### Analysis

In this part of the analysis, our objective is to identify suspicious expenses based on the total amount of money spent using classification models. The classification models used for this analysis include, Naive Bayesian Classifier, K-Nearest Neighbour (KNN), and AdaBoost. From the dataset, a new column was created to categorize expenses that exceeded $2000 as a “high expense”, and this threshold was used as the target variable for classification. Prior to this, we attempted to instead identify suspicious transactions which exceeded the average value calculated for the expense, but it only yielded 46 results. Training on these results was attempted, but the model was completely unable to identify suspicious transactions. Therefore, we decided to shift to identifying expensive transactions, which still allowed us to flag potential cases for further investigation. The $2000 threshold was chosen with the goal of having enough data identified as a high expense in order for the training to be successful.

Figure 12: Performance Metric of Naïve Bayes

A screenshot of a computer screen

Description automatically generated

Note: Own Work

Figure 13: Performance Metric of K-Nearest Neighbours

A screenshot of a computer screen

Description automatically generated

Note: Own Work

Figure 14: Performance Metric of AdaBoost

A screenshot of a computer screen

Description automatically generated

Note: Own Work

### Classification Results

Based on the performance metrics of each classifier model, K-Nearest Neighbours (KNN) seemed the most suitable for our purpose. KNN offered a balanced performance in capturing true positives and avoiding false negatives. The objective is to predict high expense items and identify areas where more money was being spent than necessary. For this context we want to prioritize recall because we would rather have false positives than false negatives, so they can be investigated, and KNN has the highest recall. Although Naïve Bayes has a high precision, meaning it made fewer mistakes in positive predictions, it has a lower recall and F1 score, indicating it does not capture actual positive cases as well as KNN did. The same can be seen with AdaBoost where it also has high accuracy and precision but the recall and F1 score are low in comparison to KNN.

## REFERENCES

[1] Governemnt of Canada. (2024, July, 19). Proactive Disclosure - Travel Expenses[Data set]. Treasury Board of Canada Secretariat. https://open.canada.ca/data/en/dataset/009f9a49-c2d9-4d29-a6d4-1a228da335ce  
   
[2] Governemnt of Canada. (2024, July, 19). Proactive Disclosure - Hospitality Expenses[Data set]. Treasury Board of Canada Secretariat. https://open.canada.ca/data/en/dataset/b9f51ef4-4605-4ef2-8231-62a2edda1b54

[3] Secretariat, T. B. of C., & Open Government, T. B. S. of C. (n.d.). *Open Government.* Retrieved August 10, 2024, from <http://open.canada.ca/en>

[4] Open Government - Government of Canada. (2013, April 26). Open Government Licence - Canada. Retrieved from http://open.canada.ca/en/open-government-licence-canada