Government Accountability Through Data Mining

By: Sean Kuehl, Mason Ward, Dominic Pham

Please raise your hand if you pay taxes

Please keep your hand raised if you think it’s all spent perfectly

Purpose

The goal of our project is to analyze government expenses in order to find significant trends and patterns. We will then use these trends and patterns to create tools and recommendations to improve efficiency and reduce waste and corruption in government expenditure.

Data

For this project we selected two Open Data datasets. One contains travel expenses by government employees, and the other contains hospitality expenses made by government employees.

-Open Data

-Discretionary Expenses

Why?

Highlights

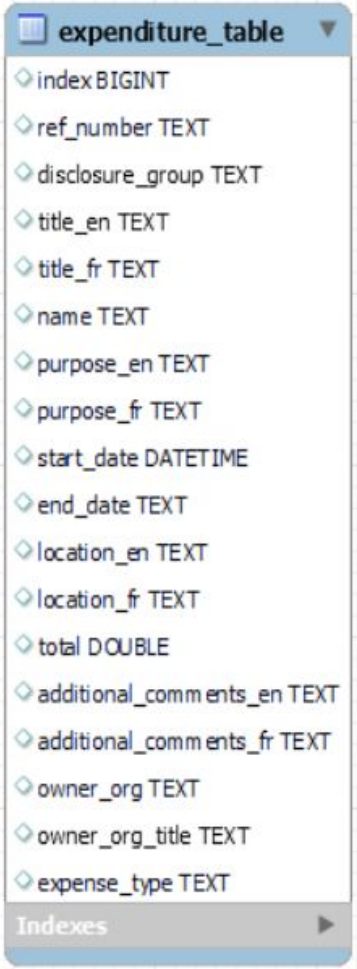
Hospitality Expenses: Board Meeting, Total Cost: $7,456.02

Travel Expenses: Operational Activities, Total Cost: $4,173.35

There may be room for criticism

Data

The Data contains several useful fields with which we can look for patterns



Pre-Processing

-Fix Formatting of Name and Location Columns

-Add “Expense Type”

-Combine Into Data Warehouse On Common Columns

-This Is Official Government Reporting?

Data Cleaning

-Remove Duplicates

-Impute Mean Into Numeric Columns

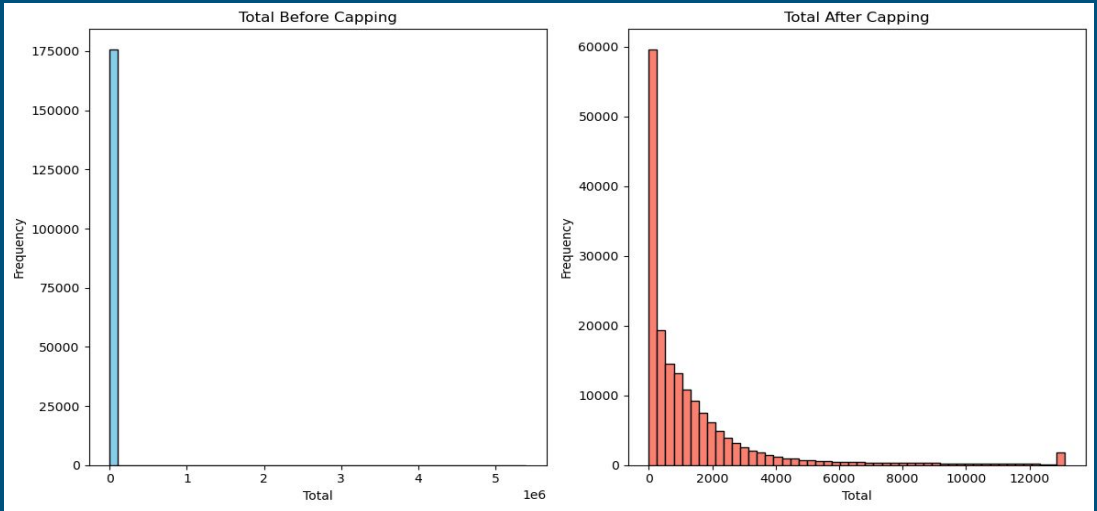
-Impute Mode Into Categorical Columns

-Impute “Unknown” Into Missing Names for Better Context

Outliers and Binning

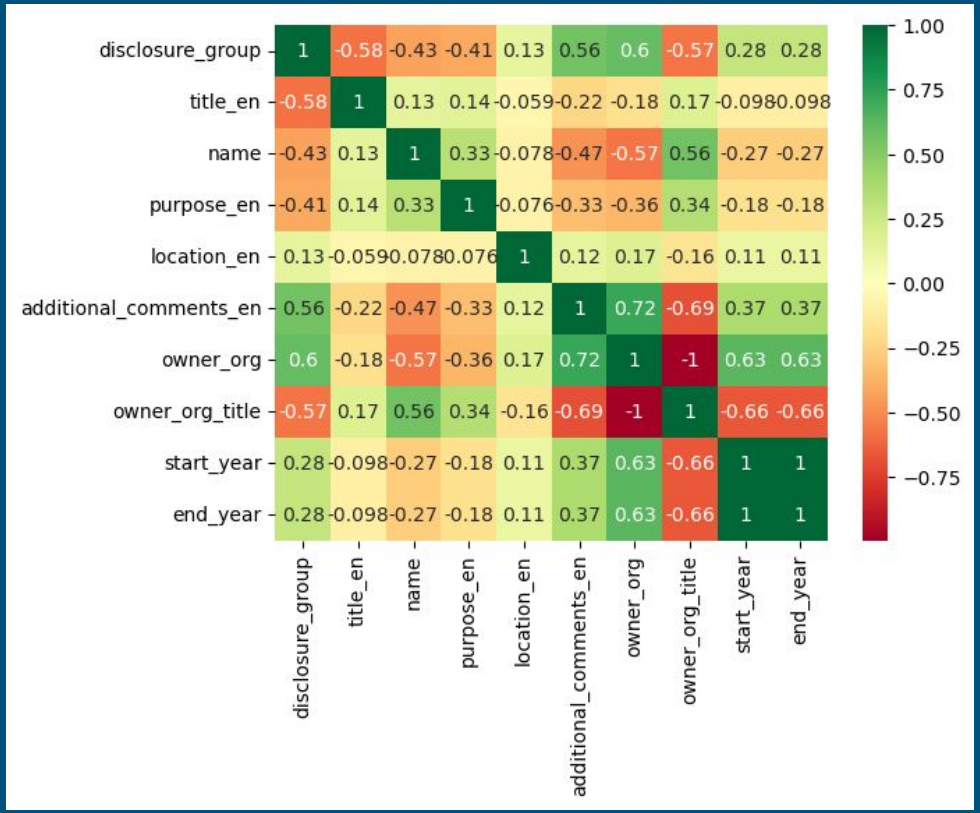
-Z-Score, IQR and Isolation Forest Method Outputs Combined, Outliers Capped

-Integer Encoding applied and “Low”, “Medium”, “High” ranges used for binning



Cluster Analysis: Correlation Analysis

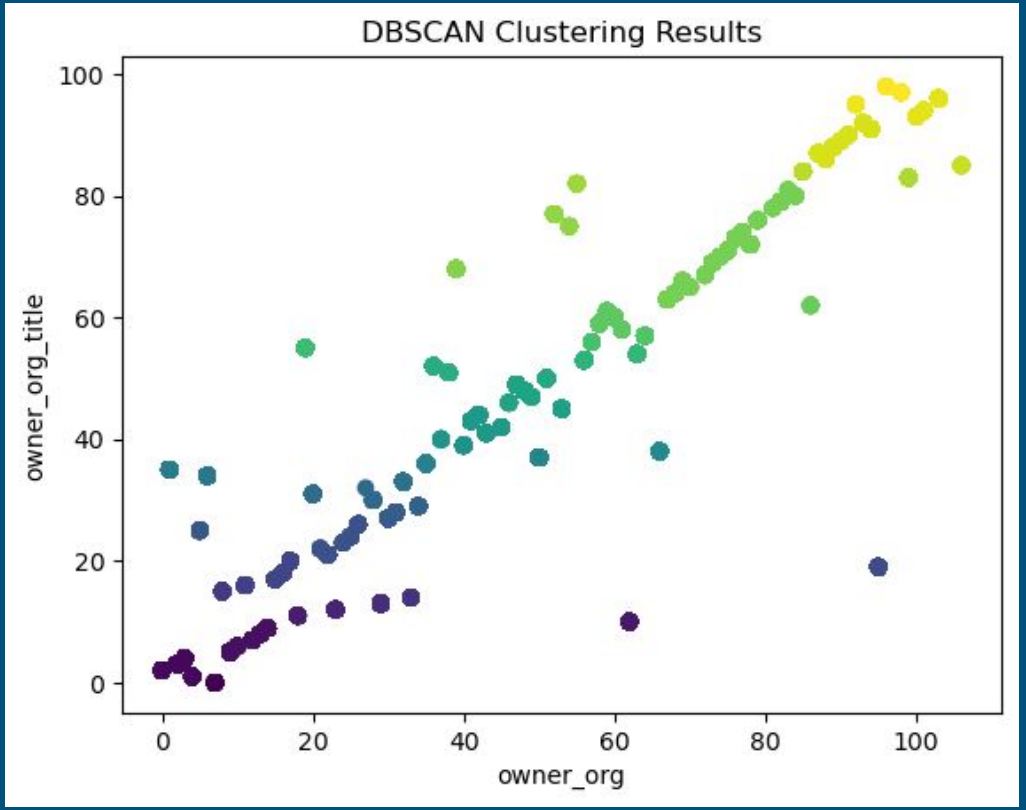
-Are There Meaningful Clusters In This Data?



Cluster Analysis: Group Discovery

-Looks Like Certain Positions Within Certain Organizations Form Groups

-What Can This Tell Us About Expenses?

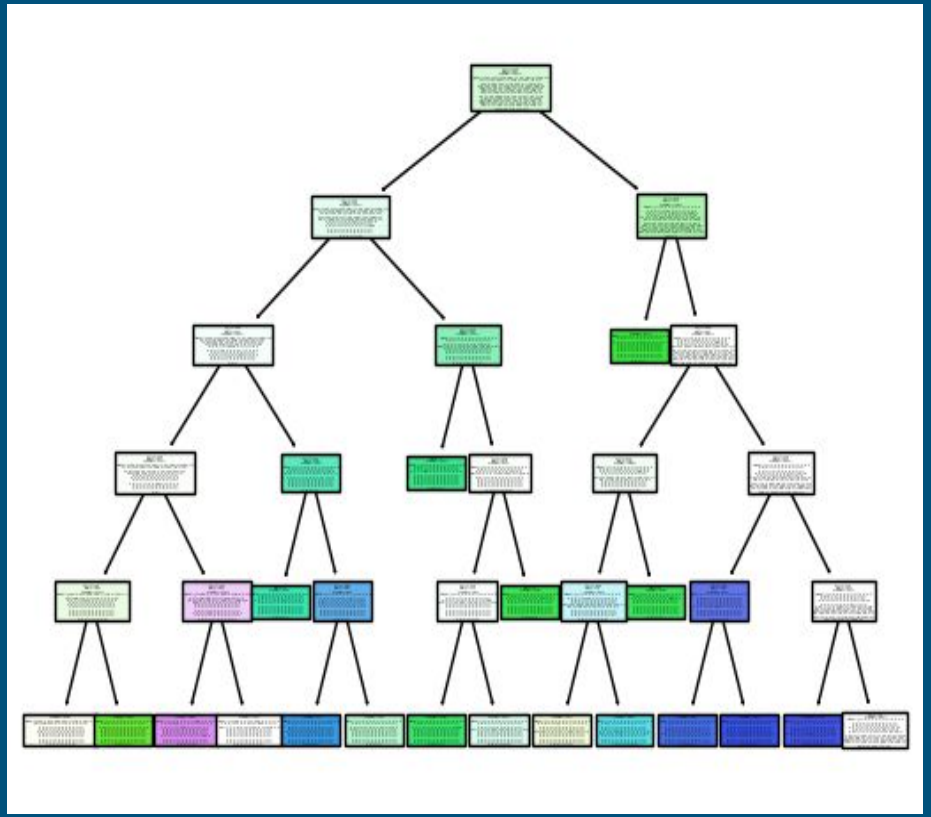


Decision Tree Prediction

-Can We Make Predictions Based On This Data?

-Restricted To Max Depth of 5

-Approximately 70% accuracy



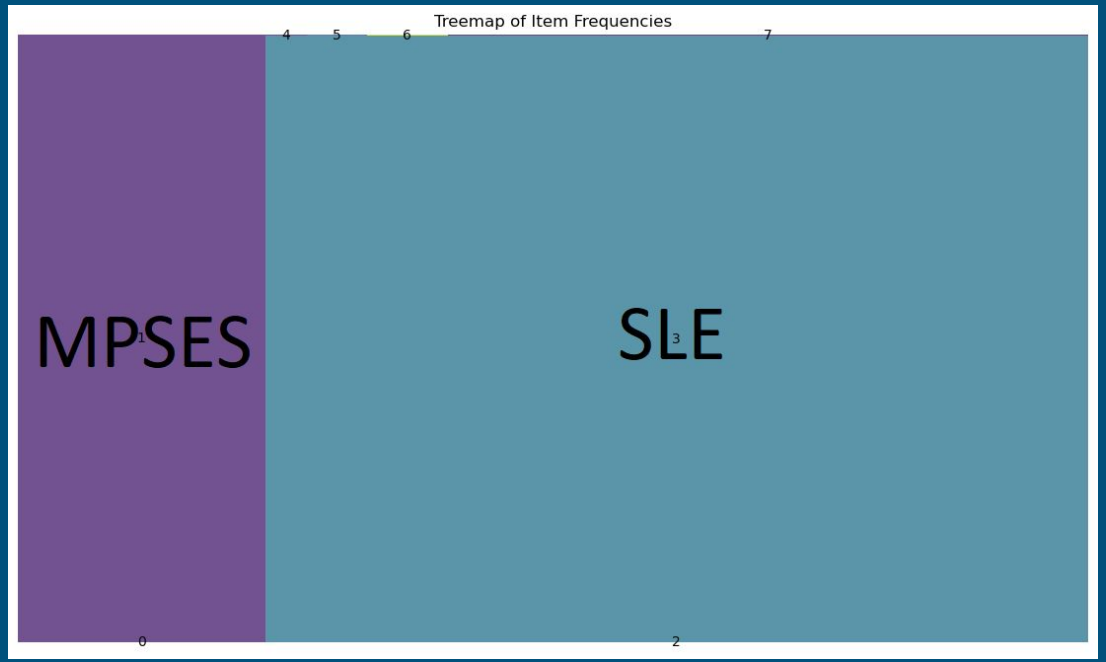
Frequency Items and Association Rules

-Are The Frequent Items In The Dataset Meaningful?

-Do The Association Rules Tell Us Anything Useful?

Frequent Items

-Some Disclosure Groups Are Much More Common Than Others



Frequent Items: Explanation

-MPSES: Minister/Ministerial adviser/Ministerial staff/Parliamentary Secretary/Exempt Staff

-SLE: Senior Level Executives

Association Rules

[17]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
23	(owner_org_title_52)	(disclosure_group_3)	0.282079	0.767287	0.235668	0.835467	1.088858	0.019232	1.414382	0.113671
24	(owner_org_title_68)	(disclosure_group_3)	0.036879	0.767287	0.032698	0.886620	1.155526	0.004401	2.052507	0.139747
18	(owner_org_title_19)	(disclosure_group_3)	0.029538	0.767287	0.029538	1.000000	1.303293	0.006874	inf	0.239796
17	(owner_org_title_13)	(disclosure_group_3)	0.033388	0.767287	0.028317	0.848112	1.105338	0.002699	1.532134	0.098591
27	(owner_org_title_71)	(disclosure_group_3)	0.021032	0.767287	0.021032	1.000000	1.303293	0.004894	inf	0.237712
29	(owner_org_title_8)	(disclosure_group_3)	0.020912	0.767287	0.020171	0.964539	1.257077	0.004125	6.562498	0.208872
5	(owner_org_title_59)	(disclosure_group_1)	0.032025	0.230830	0.019287	0.602244	2.609036	0.011894	1.933775	0.637120
13	(owner_org_title_123)	(disclosure_group_3)	0.018608	0.767287	0.016263	0.874004	1.139083	0.001986	1.846980	0.124416
14	(owner_org_title_126)	(disclosure_group_3)	0.018733	0.767287	0.014951	0.798112	1.040174	0.000577	1.152682	0.039359
20	(owner_org_title_3)	(disclosure_group_3)	0.017330	0.767287	0.014444	0.833443	1.086220	0.001146	1.397196	0.080776
31	(owner_org_title_91)	(disclosure_group_3)	0.016092	0.767287	0.013748	0.854307	1.113412	0.001400	1.597281	0.103526
10	(owner_org_title_108)	(disclosure_group_3)	0.013639	0.767287	0.013639	1.000000	1.303293	0.003174	inf	0.235931
6	(owner_org_title_83)	(disclosure_group_1)	0.011820	0.230830	0.011820	1.000000	4.332188	0.009091	inf	0.778370
8	(owner_org_title_103)	(disclosure_group_3)	0.010850	0.767287	0.010759	0.991588	1.292329	0.002434	27.663721	0.228685

Association Rules: Explanation

(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))

Classification Model

-Can We Put Everything Together and Make Meaningful Predictions On Expenses?

-Can These Predictions Be Used To Help Improve Accountability?

Classification Model

-K-Nearest Neighbours Model

-Predict High Expense Transactions

-81% Accuracy

-Not Highest Precision, but Better Recall and F1

-Better A False Positive Than Negative

Results

-Our Analysis Has Shown That There Are Meaningful Trends and Relationships In the Government’s Organizational Expenses

-These Trends and Relationships Can Be Effectively Used To Predict Important Aspects Of the Data

Recommendations

-Implement Data Entry Standard To Improve Data Integrity and Ease Future Analysis

-Use Prediction Model To Assist In Recommending Expenses To Audit

Thank You

Any Questions?