Toronto Pcard analysis Report

Objectives:

Using the city of toronto's pcard expense dataset

- Characterize the expenditures of the City's Divisions or Cost Centres by establishing the number and frequency of transactions, their typical amounts, dispersion, etc.
- Identify meaningful groups of Divisions or Cost Centres that behave similarly in terms of their expense behaviour.

Identify meaningful anomalies in the data that may require closer examination:

Dataset Details

This section summarizes the dataset that was used for this analysis.

- The dataset contains 103 excel files with each file containing pcard expenses for a given month.
- The dataset contains data for the period 2010 to 2019.
- Each file contains the following fields:
    - Division: Name of City of Toronto division
    - BATCH-TRANSACTION ID: Unique identifier assigned by credit card company
    - TRANSACTION DATE: Date of purchase
    - CARD POSTING DATE: Date purchase posted by credit card company
    - MERCHANT NAME: Name of vendor where purchase was made
    - TRANSACTION AMOUNT: Amount of purchase in Canadian currency (including amounts after conversion from other currencies)
    - TRANSACTION CURRENCY: Currency of country
    - ORIGINAL AMOUNT: Amount of Purchase in the currency of purchase (pre-exchange)
    - ORIGINAL CURRENCY: Currency in which purchase was made
    - G/L ACCOUNT: Description of the type of expense made and to be recorded
- The dataset contained the following data quality issues:
    - The column names where not identical in all files
    - 2 files contained more than 1 sheet.

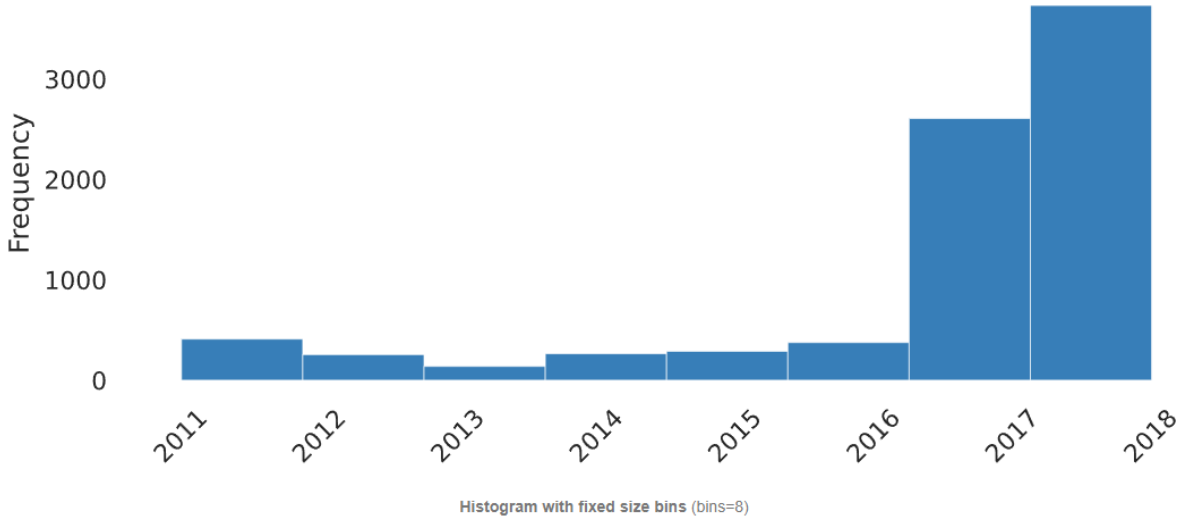
Data cleaning and preprocessing

The following data pre-processing steps where applied to the original data step:

- Fuzzy matching was used in combination with manual matching to match field names in the different files.
    - Original dataset contained 46 fields before pre-processing and 16 fields after pre-processing. See Appendix A for the results of matching
- The "Division" field also contained a lot of duplicates due to spelling errors. A combination of fuzzy and manual matching was also used to match the divisions. See Appendix B for results of

matching. The matching resulted in reducing the number of divisions from 102 to 66.

- Data scope: Total size of dataset before scoping applied - 493K, after scoping applied - 370K. 24% of the data was filtered out. The following filters were applied to the data before performing the analysis..
    - Period: Only transactions between 2011 to 2018 were considered because these years contain full years of transactions.



.

    - Only transactions with valid transaction id
    - Only transactions with a valid transaction date
    - Only transactions with a value greater than or equal to zero.
    - Only transactions with a non missing division

Feature engineering

In order to capture information from the data which can help in the analysis, new features were created. The feature creation was carried out in two steps:

1. Transaction features: These are feature created per transaction:
    a. Transaction greater than 50 CAD:Boolean flag field which is True if the transaction amount is greater than 50 CAD and false otherwise.
    b. Transaction day of week.
    c. Transaction month.
    d. Transaction year
    e. Weekday transaction: Boolean flag to check if the transaction was done within the week (Monday - Friday)
    f. Currency change: Boolean flag to check if the original currency was different from the transaction currency
    g. Difference in original and transactional amount: Boolean flag to check if the original amount is different from the transactional amount.
    h. Transaction frequency: Number of weeks since last transaction for each division.
2. Yearly features: These features were created by year for each division.
    a. Total transaction amount
    b. Average transaction amount

c. Total number of transactions: total number of batch transaction id
d. Total number of weekday transactions
e. Total number of transactions with difference in original and transactional amount
f. Total number of transactions involving a change in currency
g. Total number of cost centers
h. Total number of merchants
i. Total number of expense types
j. Total number of transactions with frequency greater than 7 days
k. Total number of transactions with transaction amount greater than 50
l. Ratio of transactions with frequency greater than 7 days
m. Ratio of transactions with transaction amount greater than 50CAD
n. Ratio of transactions involving a change in currency
o. Ratio of weekday transactions
p. Ratio of transactions with a difference in original versus transaction amount
q. Ratio of number of cost centers to number of merchants
r. Ratio of number of cost centers to number of expense types
s. Ratio of number of merchants to number of expense types
t. Evolution features were computed for features (a, b,c, d, f, h). The evolution was computed by taking the evolution over the entire period (2016 to 2018). For divisions which did not have values for these metrics in every year, we computed the evolution of the first no-zero year versus the last non-zero year.

Exploratory Data analysis

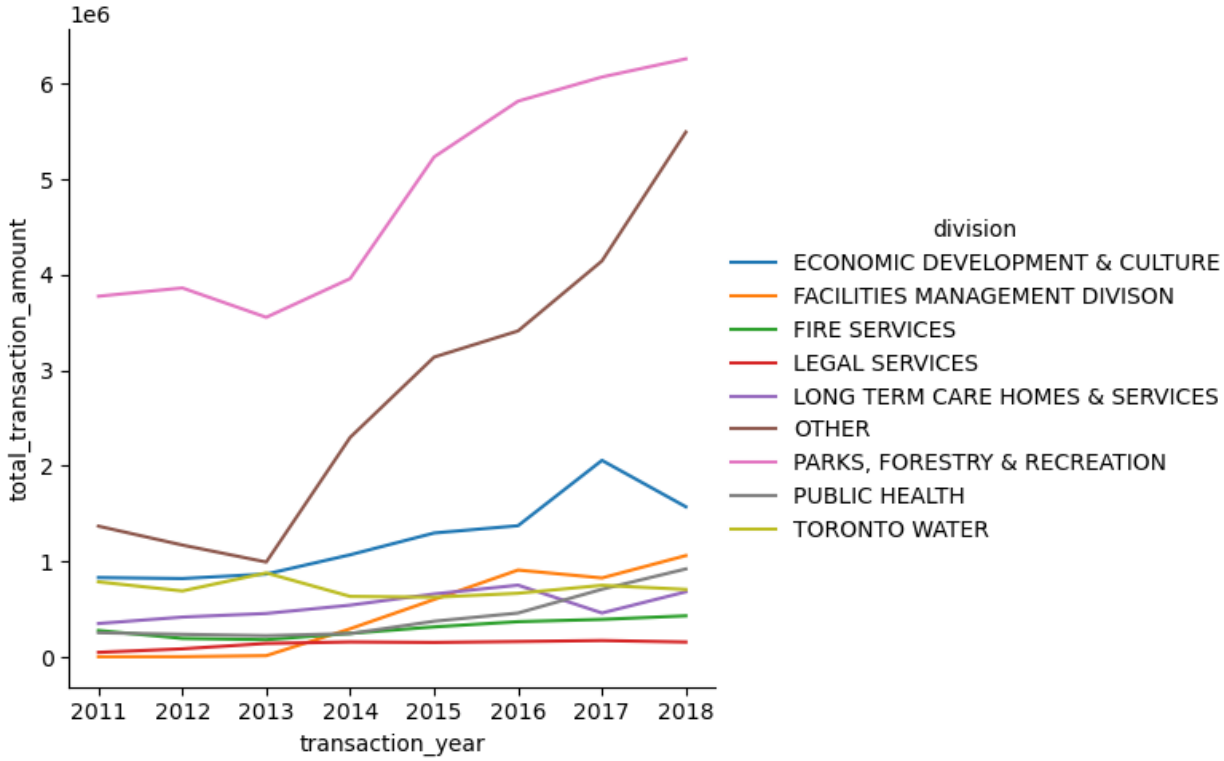
- Transactional amounts range from a couple of cents to hundreds of thousands of dollars with a majority of the transactions ranging between 30 and 220 CAD.
- A word cloud summarizing the expense types is shown below

-
- More than 80% of transactions occur during the week
- More than 90% of transactions do not involve a change in currency
- A detail report of the exploratory data analysis is attached with the source code.

Data visualization

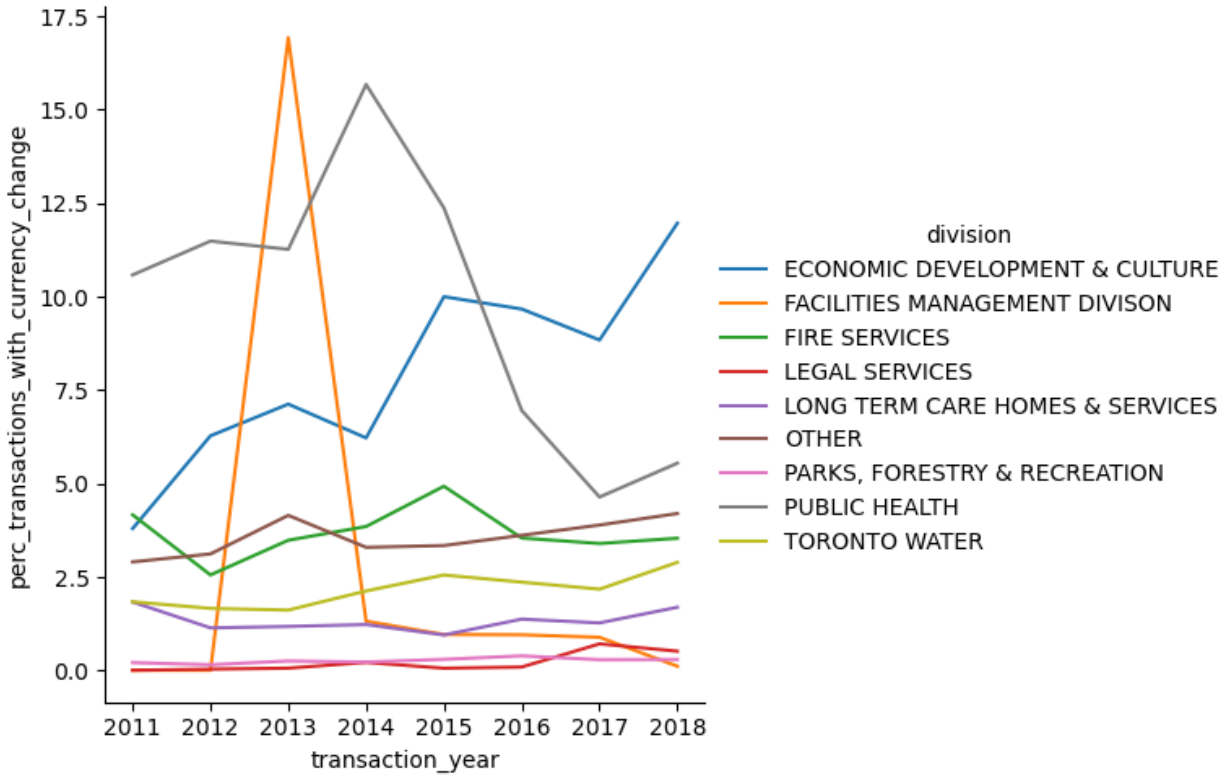
For visualization purposes, all divisions representing less than 2% of total transactions were grouped together in a new division called “OTHER”. Several visualizations were created as shown below

- Evolution of expenses by year: “PARKS, FORESTRY & RECREATION” and “OTHER” divisions show a steady increase in expenses while “LEGAL SERVICES” expenses drop significantly after 2015. “FACILITIES MANAGEMENT DIVISION” shows a sharp increase in expenses from 2013.



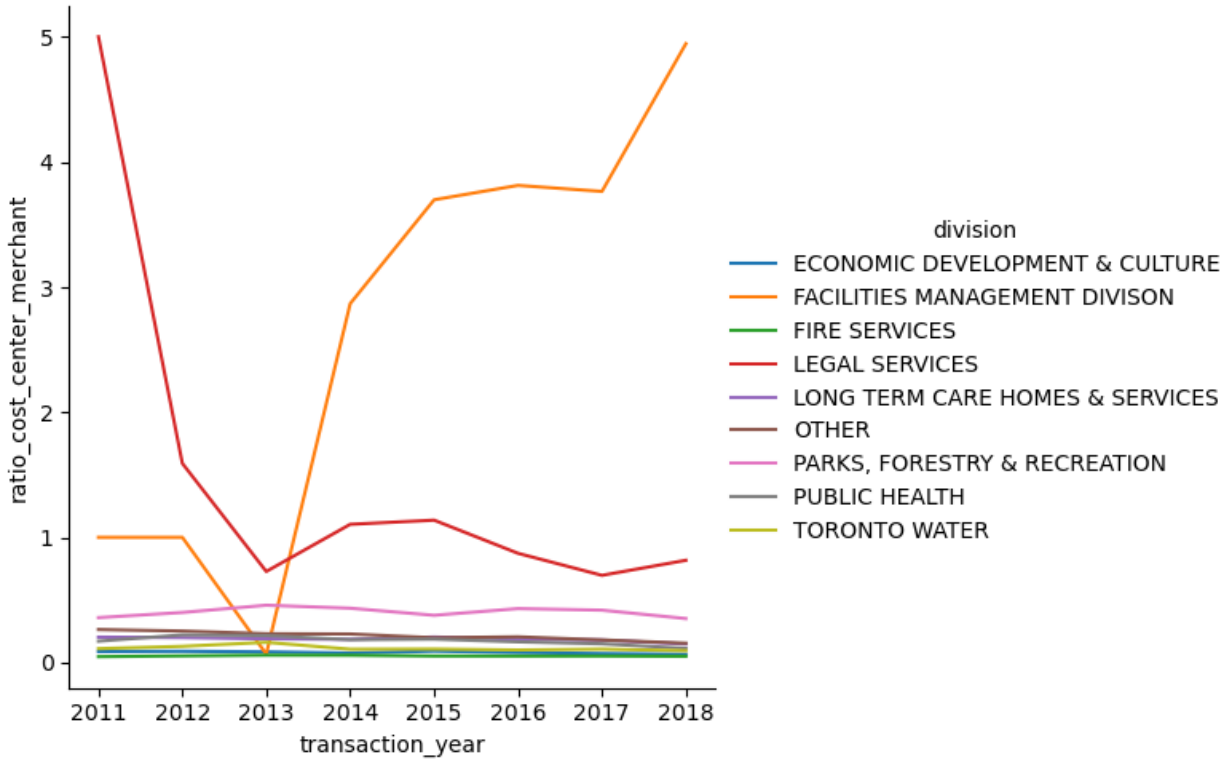
- Evolution of expenses involving a change in currency: “FACILITIES MANAGEMENT” division show a sharp rise (17.5%) in the number of transactions involving a change in currency compared to other divisions. “ECONOMIC DEVELOPMENT & CULTURE”

division show a steady rise in the number of expenses requiring a change in currency.



- Evolution of Number of cost centers to number of merchants: “FACILITIES MANAGEMENT” and “LEGAL SERVICES” show an evolution in the relationship

between number of cost centers and merchants.

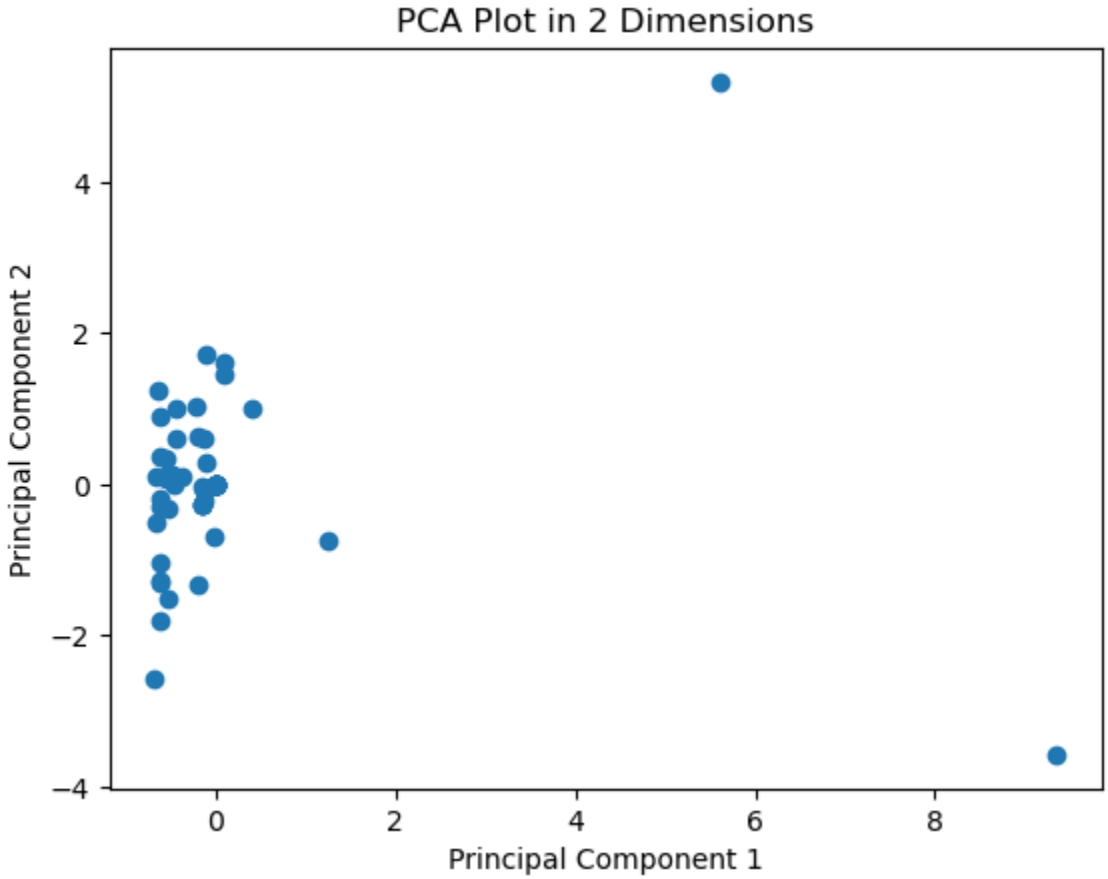


Clustering

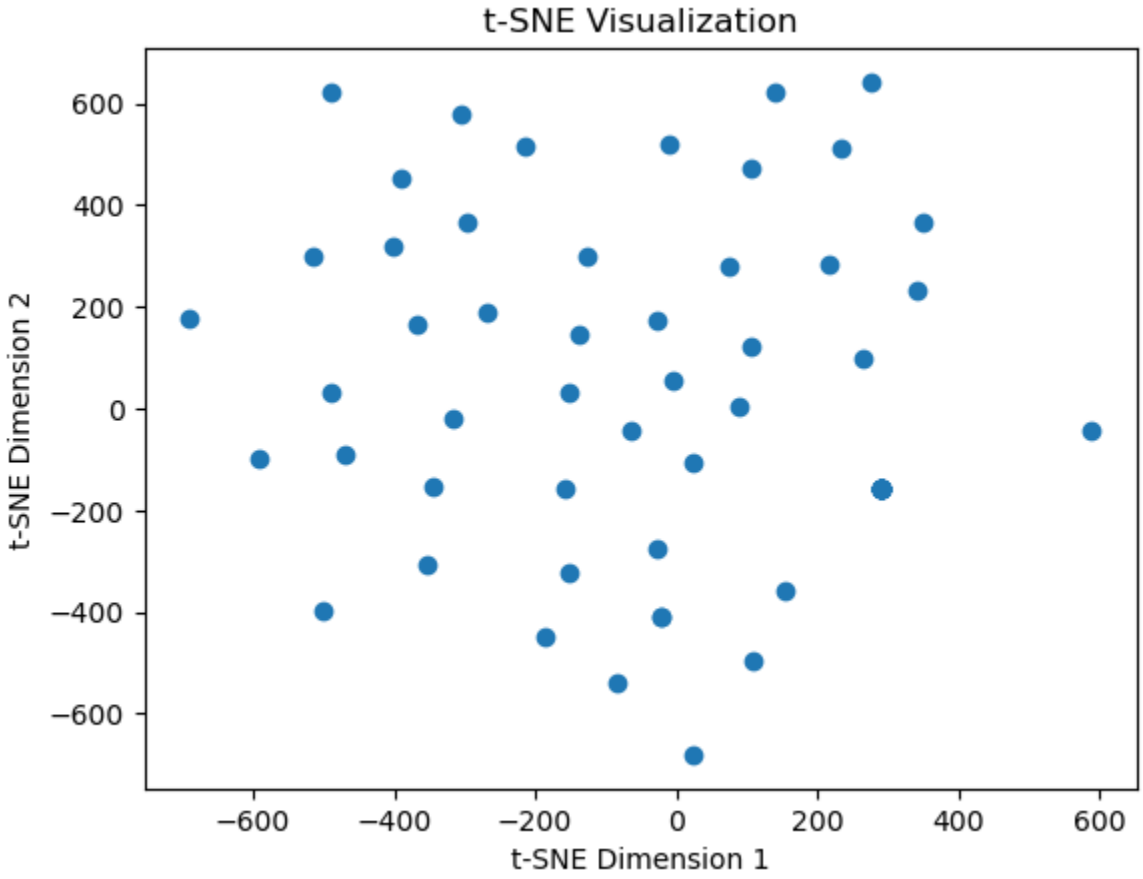
A 2 step approach was used for establishing the number of clusters.

- Variable selection: Due to the number of variables (> 30) , PCA was used to extract the top 5 variables with maximum variance. According to this approach the following variables were selected [`'total_cost_centers_2015', 'total_trans_gt_weekly_2018', 'ratio_cost_center_expense_types_2015', 'mean_transaction_amount_2012', 'total_transaction_amount_2011']`
- Performing clustering with different techniques: Data dimension reduction and visualization techniques were used to define the possible number of clusters.

- Principal component analysis (PCA): PCA analysis did not also provide much information to deduce the number of clusters

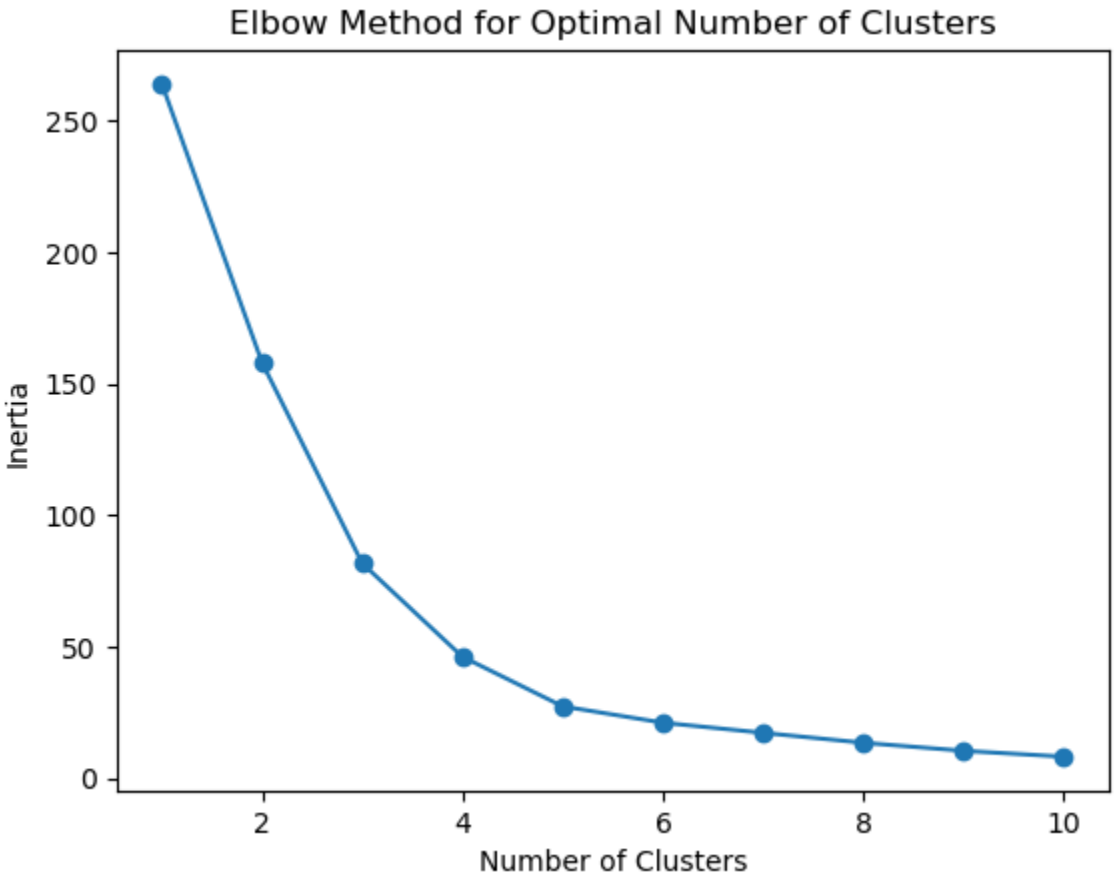


- T-distributed stochastic neighbor embedding: (t-SNE). This approach was not very conclusive as it was very difficult to deduce the number of clusters



- K-means clustering:Using the elbow method this approach suggests 3 clusters. The clusters produced were highly skewed with on cluster containing 64 out of

the 66 divisions.



- Perform clustering: Based on the above observations, HDBSCAN - Hierarchical Density-Based Spatial Clustering of Applications with Noise. This approach produced the best result among all approaches.
  - Number of clusters: 5
  - Cluster sizes:
    - Cluster 1 (21 divisions): `['SHELTER & SUPPORTIVE HOUSING ADMINISTRATION', 'ENVIRONMENT & ENERGY OFFICE', 'TORONTO BUILDING', 'EMERGENCY MEDICAL SERVICES', 'ECONOMIC DEVELOPMENT & CULTURE', 'TECHNICAL SERVICES', 'TORONTO WATER', 'LEGAL SERVICES', 'LONG TERM CARE HOMES & SERVICES', 'MUNICIPAL LICENSING & STANDARDS ', 'TRANSPORTATION', 'FIRE SERVICES', 'CITY MANAGER', 'CITY CLERKS OFFICE', 'PARKS, FORESTRY & RECREATION ', 'SOLID WASTE MANAGEMENT', 'PPFA', 'TRANSPORTATION SERVICES ', 'PUBLIC HEALTH', 'ACCOUNTING SERVICES ', 'FLEET SERVICES']`

- Cluster 2 (9 divisions): ['STRATEGIC COMMUNICATIONS', 'PENSION, PAYROLL & EMPLOYEE BENEFITS', 'OFFCIE OF THE DEPUTY CITY MGR. & CFO', 'TORONTO PARAMEDIC SERVICES', 'CHILDRENS SERVICES', 'AFFORDABLE HOUSING OFFICE', 'ENGINEERING & CONSTRUCTION SERVICES', 'CORPORATE CONTRACTS', 'SOCIAL DEVELOPMENT, FINANICE & ADMINISTRATION']
- Cluster 3 (19 divisions): ['CFO', 'CHIEF FINANCIAL OFFICER', 'PMMD', 'CHIEF INNOVATION OFFICE', 'CHIEF TRANSFORMATION OFFICE', 'OFFICE OF THE TREASURER', 'OFFICE OF THE DCM - INTERNAL SERVICES CLUSTER', 'OFFICE OF THE DCM - CORPORATE SERVICES', 'TORONTO OFFICE OF PARTNERSHIPS', "CITY MANAGER'S OFFICE", 'CORPORATE FINANCE', 'CORPORATE SECURITY', 'TREASURER', 'DEPUTY CITY MGR INTERNAL SERVICES', 'EXECUTIVE MANAGEMENT', 'EXECUTIVE MGT', 'FINANCIAL PLANNING', 'OFFICE OF THE CONTROLER', 'INTERNAL AUDIT']
- Cluster 4 (6 divisions): ['TORONTO EMPLOYMENT & SOCIAL SERVICES', '311 TORONTO', 'POLICY, PLANNING, FINANICE & ADMINISTRATION', 'CITY PLANNING', 'REVENUE SERVICES ', 'URBAN PLANNING']
- Cluster 5 (11 Divisions) : ['REAL ESTATE SERVICES', 'PURCHASING & MATERIALS MANAGEMENT ', 'STRATEGIC & CORPORATE POLICY', 'OFFICE OF EMERGENCY MANAGEMENT', 'INFORMATION & TECHNOLOGY', 'HUMAN RESOURCES ', 'FINANCIAL SERVICES', 'FINANCE & ADMINISTRATION ', 'FACILITIES MANAGEMENT DIVISON', 'COURT SERVICES', 'AfFFORDABLE HOUSING OFFICE']

- Cluster characteristics: This was obtained by looking at the summary statistics by cluster of the selected features identified in the variable selection phase (['total_cost_centers_2015', 'total_trans_gt_weekly_2018', 'ratio_cost_center_expense_types_2015', 'mean_transaction_amount_2012', 'total_transaction_amount_2011']). Appendix C contains a table with the summary of the features characterizing the different clusters.
  - Cluster1: Very high expenditure in 2011
  - Cluster 2: These are divisions relatively medium cost center to expense types in 2015

- Cluster 3: These are divisions which were were not very active before 2016
- Cluster 4: Divisions with very low cost center to expense type in 2015
- Cluster 5: These divisions are characterized by a very high number of cost centers in 2015

Anomaly detection

The probability to belong to a Cluster was used to identify divisions with anomalies based on the selected features. In practice, the features to use for clustering can be jointly decided by exploratory analysis and business expert knowledge. The top 10 anomaly divisions identified with this clustering approach include

- **SHELTER & SUPPORTIVE HOUSING ADMINISTRATION**
- **PUBLIC HEALTH**
- **FLEET SERVICES**
- **FIRE SERVICES**
- **LONG TERM CARE HOMES & SERVICES**
- **ENVIRONMENT & ENERGY OFFICE**
- **PARKS, FORESTRY & RECREATION**
- **EMERGENCY MEDICAL SERVICES**
- **ECONOMIC DEVELOPMENT & CULTURE**
- **MUNICIPAL LICENSING & STANDARDS**

Sampling approach to contrast Normal vs Anomaly division

Given that an anomaly is relative to a given cluster, a stratified random sampling approach can be used to contrast the divisions. A new variable has to be created to label the anomaly cases. This new variable together with the cluster label can be used for the different strata during the sampling process.

-

Appendix A: Result of field matching

```
{'transaction currency': 'trx currency',
 'cost centre wbs element descrption': 'cost center wbls element order description',
 'cost centre wbs element description': 'cost center wbls element order description',
 'cost centre wbs elelment description': 'cost center wbls element order description',
 'cost centrewbs element description': 'cost center wbls element order description',
 'cost centre wbs element': 'cost center wbls element order',
 'unnamed 16': 'unnamed 16',
 'card posting dt': 'card posting date',
 'card posting date': 'card posting date',
 'cost centrewbs element': 'cost center wbls element order',
 'cost centre wbs elelment': 'cost center wbls element order',
 'original currency1': 'original currency1',
 'original currency': 'original currency1',
 'cost centre wbs element order': 'cost center wbls element order',
 'cost center wbs element order': 'cost center wbls element order',
 'cost centre wbs element work order number': 'cost center wbls element order',
 'cost centre wbs element order description': 'cost center wbls element order description',
 'division': 'division',
 'divison': 'division',
 'transaction amount': 'transaction amount',
 'transaction amt': 'transaction amount',
 'batchtransaction id': 'batch transaction id',
 'batch transaction id': 'batch transaction id',
 'merchant type description': 'merchant type description',
 'cost center wbls element order description': 'cost center wbls element order description',
 'cost center wbs element order description': 'cost center wbls element order description',
 'cost centre wbs element order decription': 'cost center wbls element order description',
 'gl account description': 'gl account description',
 'gl account decription': 'gl account description',
 'gl account discription': 'gl account description',
 'transaction date': 'transaction date',
 'original amount': 'original amount',
 'merchant type mcc': 'merchant type mcc',
 'merchant type': 'merchant type mcc',
 'gl expense description': 'gl account description',
 'trxcurrency': 'trx currency',
 'trx currency': 'trx currency',
 'tr currency': 'trx currency',
 'exp type desc': 'gl account description',
 'gl account': 'gl account',
 'long text': 'gl account description',
 'cost centrewbs elementwork order number description': 'cost center wbls element order description',
 'cost centrewbs element discriprion': 'cost center wbls element order description',
 'purpose': 'purpose',
 'merchant name': 'merchant name',
 'unnamed 17': 'unnamed 17'}
```

Appendix B: Result of division matching

```
{'tranportation': 'transportation',
 'transportation': 'transportation',
 'executive management': 'executive management',
 'office partnerships': 'office partnerships',
 'water': 'water',
 'chief transformation office': 'chief transformation office',
 'environment energy office': 'environment energy office',
 'environment office': 'environment energy office',
 'environment energy': 'environment energy office',
 'pension payroll employee benefits': 'pension payroll employee benefits',
 'treasurer': 'treasurer',
 'internal audit': 'internal audit',
 'human resources': 'human resources',
 'chief financial officer': 'chief financial officer',
 'ppfa': 'ppfa',
 'social development finance administration': 'social development finanice administration',
 'social development finanice administration': 'social development finanice administration',
 'deputy city mgr internal services': 'deputy city mgr internal services',
 'parks forestry recreation': 'parks forestry recreation',
 'emergency medical services': 'emergency medical services',
 'afffordable housing office': 'afffordable housing office',
 'affordable housing office': 'affordable housing office',
 'chief innovation office': 'chief innovation office',
 'employment social services': 'employment social services',
 'fire services': 'fire services',
 'urban planning': 'urban planning',
 'fleet services': 'fleet services',
 'shelter supportive housing administration': 'shelter supportive housing administration',
 'shelter support housing administration': 'shelter supportive housing administration',
 'shelter support housing adminstration': 'shelter supportive housing administration',
 'engineering construction services': 'engineering construction services',
 'strategic communications': 'strategic communications',
 'city planning': 'city planning',
 'transportation services': 'transportation services',
 'long term care homes': 'long term care homes services',
 'long term care homes services': 'long term care homes services',
 'executive mgt': 'executive mgt',
 'policy planning finance administration': 'policy planning finanice administration',
 'policy planning finanice administration': 'policy planning finanice administration',
 'corporate security': 'corporate security',
 'legal services': 'legal services',
 'revenue services': 'revenue services',
 'corporate finance': 'corporate finance',
 'offcie deputy city mgr cfo': 'offcie deputy city mgr cfo',
 'deputy city mgr cfo': 'offcie deputy city mgr cfo',
 'economic development culture': 'economic development culture',
 'office dcm internal services cluster': 'office dcm internal services cluster',
 'office dcm corporate services': 'office dcm corporate services',
 'office controler': 'office controler',
```

Appendix C: Cluster features

	total_cost_centers_2015	total_trans_gt_weekly_2018	ratio_cost_center_expense_types_2015	mean_transaction_amount_2012	total_transaction_amount_2011
	mean	mean	mean	mean	mean
cluster					
-1	96.529412	0.0	1.300189	323.797692	402603.719474
0	7.333333	0.0	2.512372	333.474123	NaN
1	NaN	0.0	NaN	NaN	NaN
2	11.750000	0.0	0.620916	318.219919	5131.856667
3	291.500000	0.0	9.159722	301.372559	885.353636