CAT Digital Team × P

TEAM MEMBERS: Yury A. Kuleshov, Pum Khai, Leying Zhou (Data Mine); Girija Bhamidipati, Thrishna Bhandari, Haleigh Gronwold, Pawandeep Kaur, Rthvik Raviprakash, Kayla Seeley (Capstone)
MENTORS: Dr. Davide Gerbaudo, Dr. Sridhar Ramaswamy; Kevin R. Sellers (CAT); Dr. Jean F. Honorio Carrillo (Capstone)
TAs: Jayanth Krishnamurthy (Data Mine), Tanmaya Udupa (Capstone)

INTRODUCTION

- **About Caterpillar & CAT Digital**
- **Cat Digital** is a digital and technology business unit within Caterpillar Inc.
  - We're focused on using data, technology, advanced analytics and AI capabilities to help our customers build a better world.
  - Cat Digital embodies the bits-and-bytes coming from pieces of machinery, customers and dealers, and is charged with ensuring data flows smoothly and safely through our platform and to applications.

Caterpillar Inc. is the world's leading manufacturer of construction and mining equipment, off-highway diesel and natural gas engines, industrial gas turbines and diesel-electric locomotives.

For nearly 100 years, we've been helping customers build a better, more sustainable world and are committed and contributing to a reduced-carbon future. Our innovative products and services, backed by our global dealer network, provide exceptional value that helps customers succeed.

OUR SOLUTIONS HELP OUR CUSTOMERS BUILD A BETTER WORLD.

107,700 Global Full-Time Employees

62% Sales and Revenues Outside the United States

2021 Sales and Revenues **$51B**

PRODUCT LINE

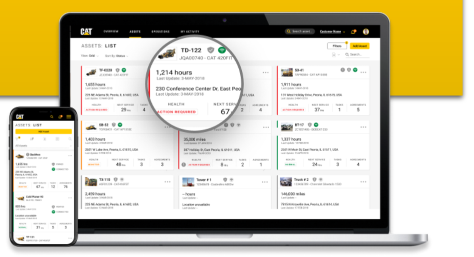
4M+ Cat® Products at Work Around the World*

*Includes discontinued products

Construction Equipment | Mining Equipment | Off-highway Diesel & Natural Gas Engines | Industrial Gas Turbines | Diesel-Electric Locomotives

- **Motivation & Problem**
  Currently, Caterpillar is using a clustering model based primarily on industry- and company-based rules to categorize invoices and parts. This strategy is effective for bootstrapping and is easy to understand and explain to their business partners. However, these rules are not necessarily backed by any scientific processes and trends/rules can change often. Hence, Caterpillar needs a more robust solution with proper validation to classify and categorize invoices from parts sales to customers.

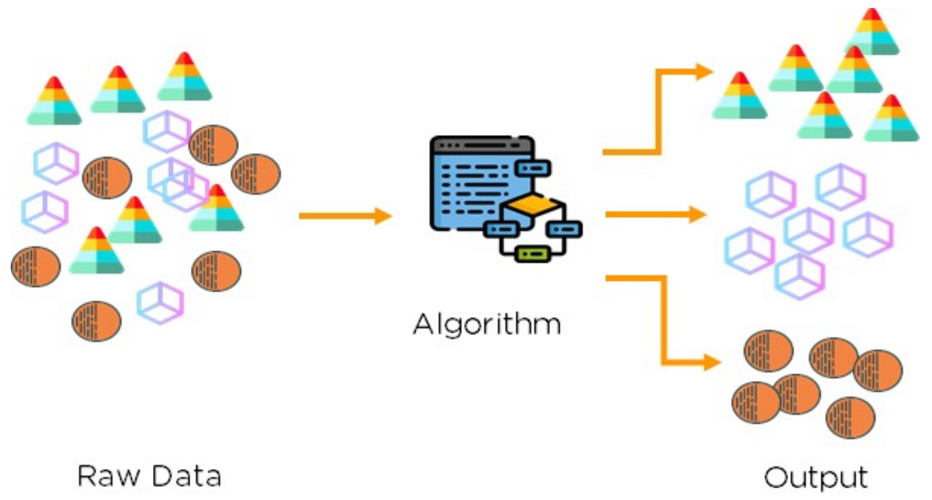
- **Objective**
  The goal of this project is to explore a ***novel document clustering algorithm*** to be used with invoices. The new algorithm should leverage a similarity metric built upon the numerical and categorical features associated with the line items listed on the invoice documents.



RESEARCH METHODOLOGY

- **Research Design Ideas**
  Since we want to better cluster the invoices, we are trying to determine ***distance metrics*** for this type of data specifically. → ***The similarity metric*** should work with the "invoice" data type, namely a mix of numerical, categorical, and textual data. → ***"Similarity metric"*** feature that given two invoices it can determine how similar or dissimilar they are.

- **Key Approaches**
  - Similarity metrics
  - Tf-IDF clustering
  - Natural Language Processing



Stage 1: Looking for pattern via correlation → Stage 2: Searching for appropriate clustering algorithm → Stage 3: Implement the algorithm into an API in python → Substage 3-1: Model Building → Substage 3-2: Testing & Unit testing → Stage 4: Document for deliverable

KEY FINDINGS & SOLUTIONS

- **Data Description & Data Attributes**

```
Column descriptions.
Columns with a HASH() have been obfuscated for data confidentiality

  HASH(INVOICE_ID) -- primary key to identify an invoice
, HASH(DEALERCODE) -- code to identify the dealership where the part invoice was issued
, INVOICE_DATE  -- date at which the invoice was issued
, SALES_METHOD -- how the parts were sold: OTC= over the counter, WO= work order with the repair performed at the dealership, B2B = business-to-business internet purchase
, HASH(SERIALNUMBER) -- serial number (more precisely its prefix) of the machine(s) being repaired with the parts in this invoice
, HASH(SALES_MODEL) -- sales model of the machine(s) being repaired.
, HASH(PART_NUMBER) -- code to identify a replacement part
, PART_NAME -- part name as given by the engineer (can be very specific/generic, have typos, abbr. etc.)
, PART_IDENTIFIER -- standardized generic description of the part
, HASH(SOURCE_OF_SUPPLY) -- whether the part is supplied by CAT (majority) or by another vendor
, HASH(PART_PRICING_CODE) -- category of parts used to determine its price
, HASH(MAJOR_CLASS) -- coarse category of parts
, HASH(COMMERCIAL_GROUP_TYPE) -- intermediate category of parts
, HASH(AFTERMARKET_OFFER) -- fine category of parts
, MACHINE_COMPONENT_L1  -- coarse component where this part is installed
, MACHINE_COMPONENT_L2  -- intermediate component where this part was installed
, MACHINE_COMPONENT_L3  -- fine component where this part was installed
, PART_QUANTITY -- quantity of this line item
```

- **Data Cleaning Principles**
  - Reduce the noise from negative price and quantity
  - Clarify electric machine component numbers
  - Improve null values
  - Delete non-Caterpillar parts in the dataset
  - Create dummy variables

- **Data Construction**

```
sample.groupby(['invoice_id', 'major_class']).agg(
  usd_by_maj=pd.NamedAgg(column='price', aggfunc='sum'),
  n_pns_by_maj=pd.NamedAgg(column='part_number', aggfunc='nunique')
)
```

  - Group by invoice_ids
  - Group by major_class
  - Carry out stratified sampling

- **Data Format**
  - Create Sales_method as dummy variables

- **Models**
  - DBSCAN model (Figure 2)
  - K-Means with 3 clusters (Figure 3)
  - Hierarchical clustering (Figure 4)

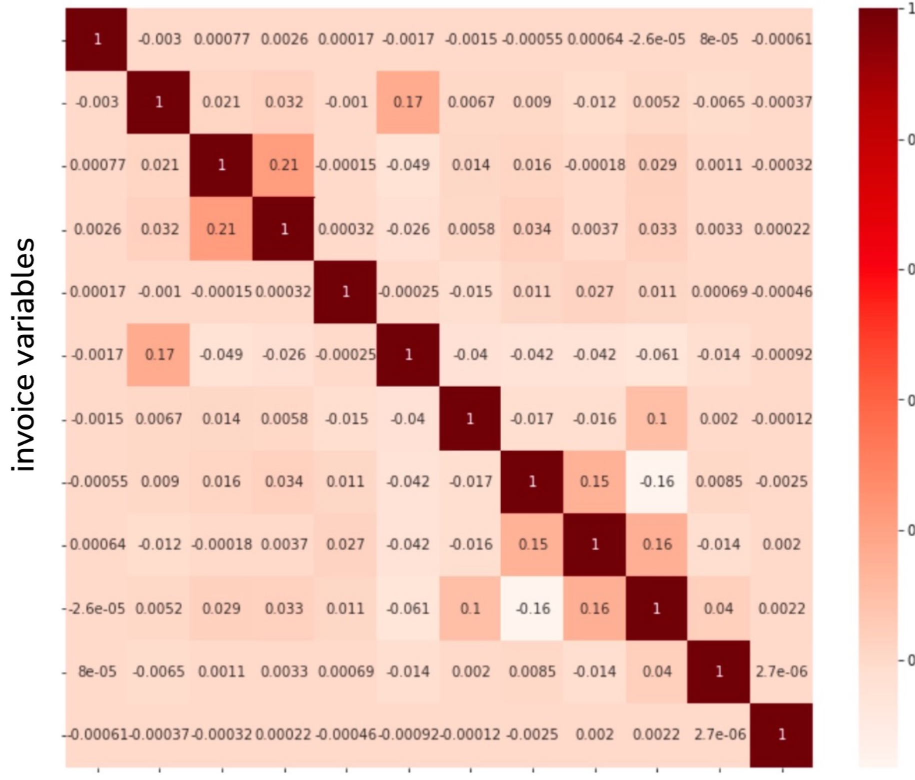


Figure 1. Heatmap (from 10,000 rows)
- No significant correlations between attributes
  - most numerical attributes are hashed
  - others are categorical
- Sales_model & sales_method are two specific attributes of relevance for data mining and business purposes since they provide insights into how big/small the repair was and the method in which it was accomplished

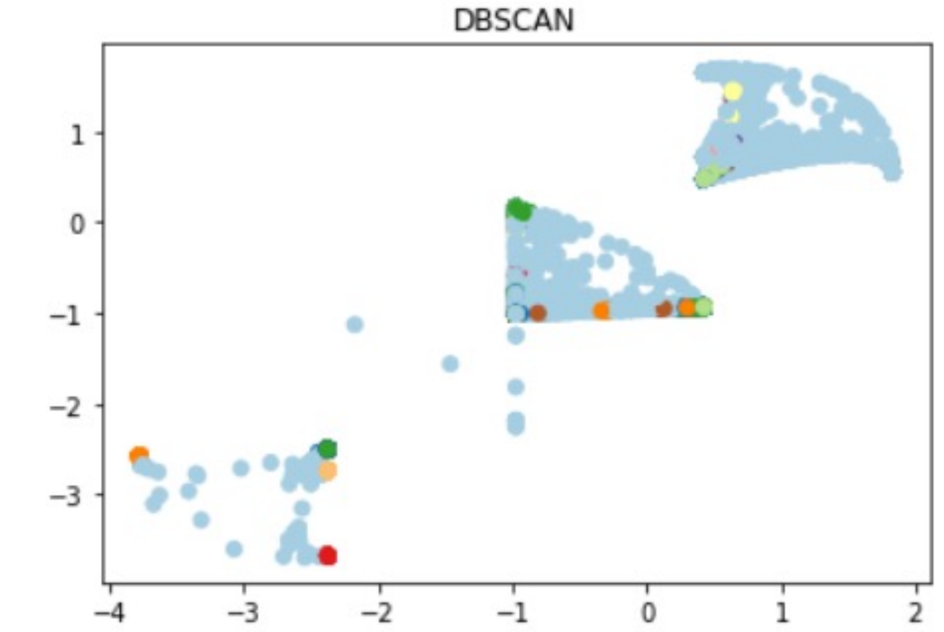


Figure 2. DBSCAN model
- Optimal number of clusters to be 62

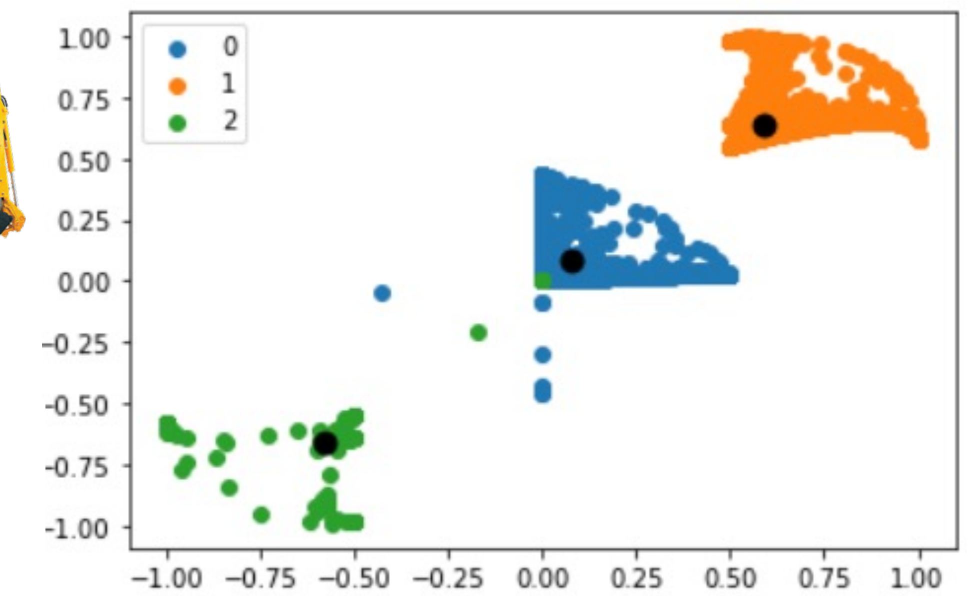


Figure 3. K-Means with 3 clusters
- Chosen based off an elbow curve

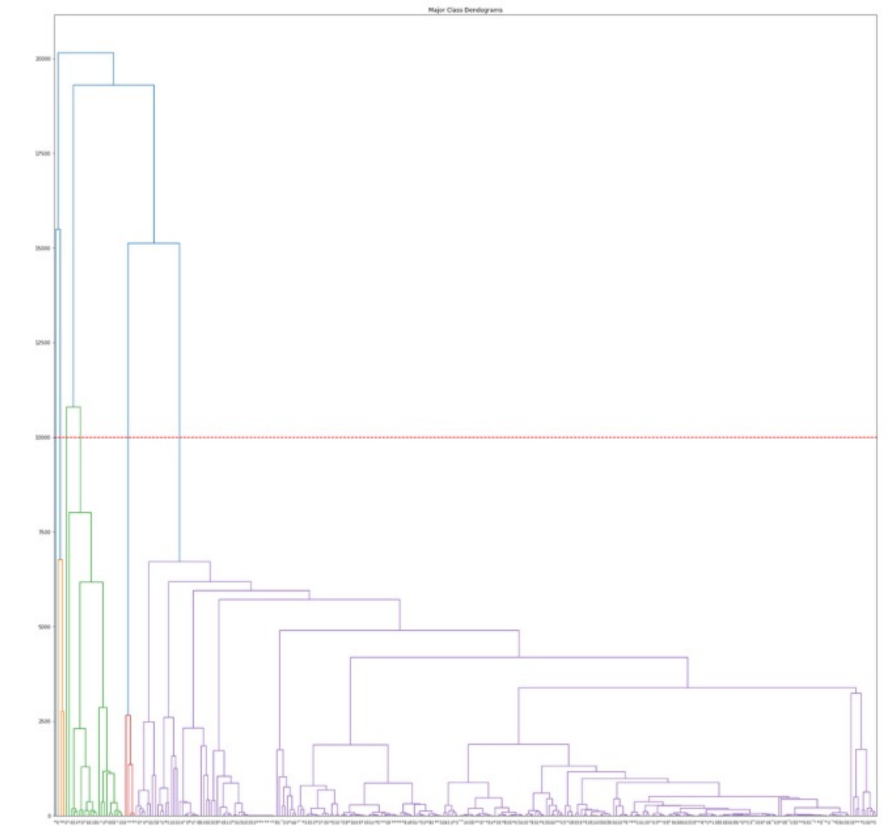
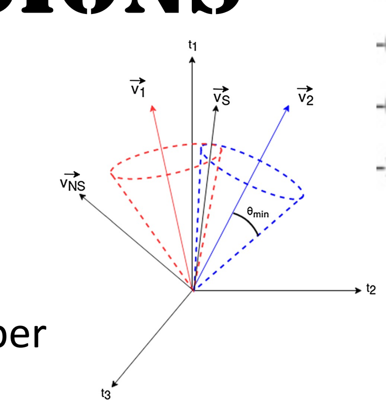


Figure 4. Hierarchical clustering
- Dendrogram produced using 6 clusters

CONCLUSIONS

- **Improvement of clustering algorithm**
  - **Cosine similarity** is more useful than Euclidean distance.
  - Principal Component Analysis did not seem to improve the models much in reducing the number of dimensions.
  - To evaluate, we have been looking at the number of data points in each cluster to see how evenly they are spread and also looking at inter-cluster distance.



FUTURE GOALS

- Include more advanced clustering ideas such as Gower Distance, and k-medoids, etc.
- Further, to improve upon the algorithms/model with a computationally efficient metric in a way that millions of invoices can be processed by computers compared to traditional document clustering algorithms.