Determining How Americans Get Hurt On the Job

Georgetown University Certificate in Data Science, Summer 2019

GEORGETOWN UNIVERSITY

Agenda & Team Introductions

Today's Agenda

1. ***Agenda & Team introduction***
2. ***Project context***
3. ***Methodology***
4. ***Results & Reflections***

20 Minutes

6. ***Feedback, reactions, questions***

5 Minutes

Introducing….

Project Context

Background

Diagnosis/procedure code systems are used in multiple industries for:

- Documenting care (electronic health records)
- Billing for care (insurance claims)
- Research and analysis, surveillance

International Classification of Diseases (ICD-10-CM):

- 70,000 diagnosis codes, 60,000 procedure codes
- Great detail -- but too much information for most purposes.

Less is More

"Groupers" are analytic tools that aggregate sets of similar codes into broad, clinically meaningful categories.

Example: "Clinical Classification Software" groups 14,000 codes into 275 categories. Much more manageable!

Traditionally, groupers have been very expensive to develop.

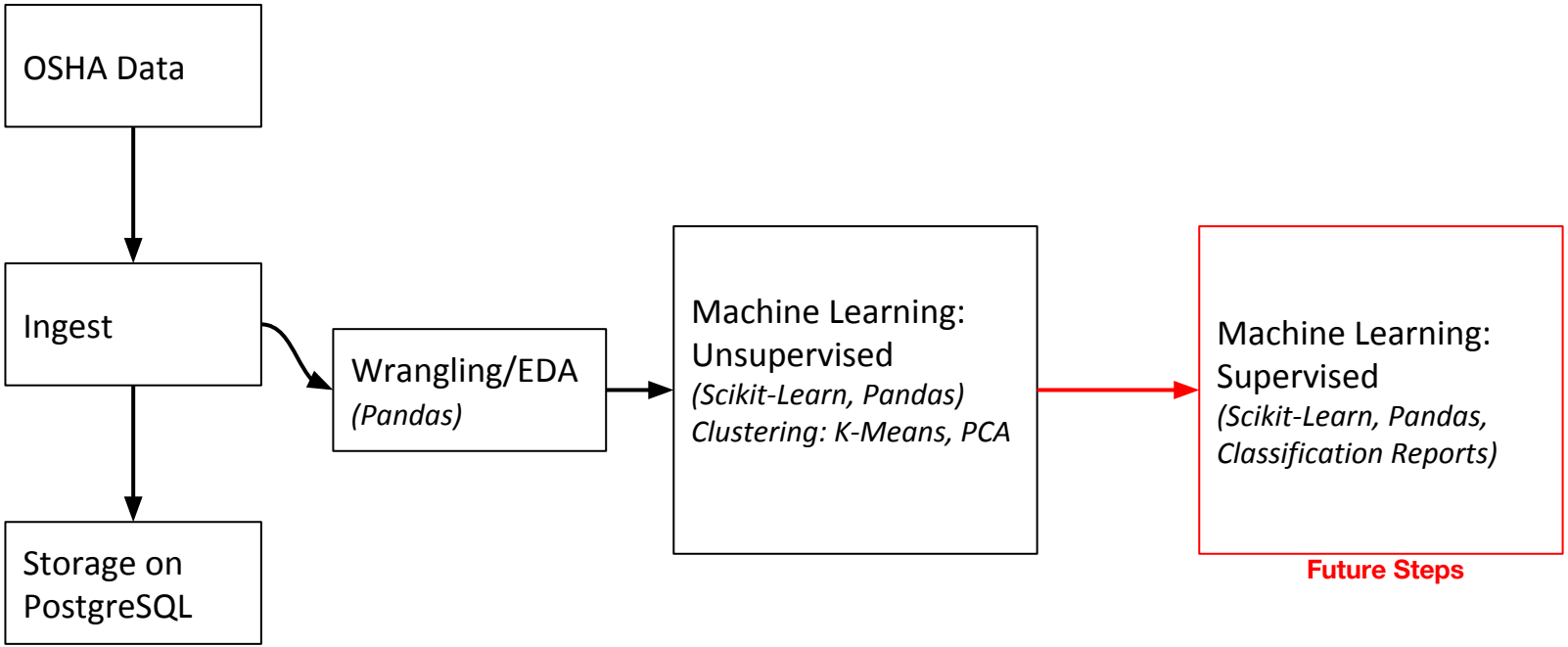
Grouping == Clustering

Objective: To evaluate the feasibility of using k means clustering to identify clinically meaningful groups of individual injury codes.

Question: Could k means clustering produce a preliminary draft of groups, which could then be finalized by expert reviewers?

Methodology

Data Pipeline

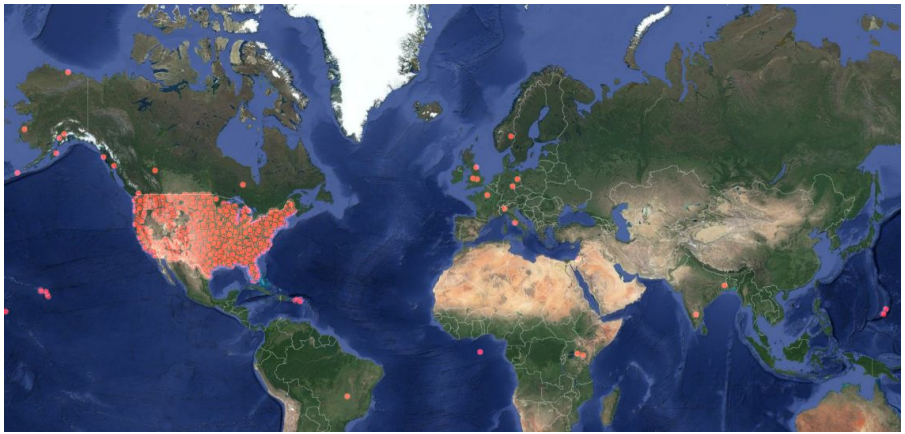


Ingestion & Wrangling

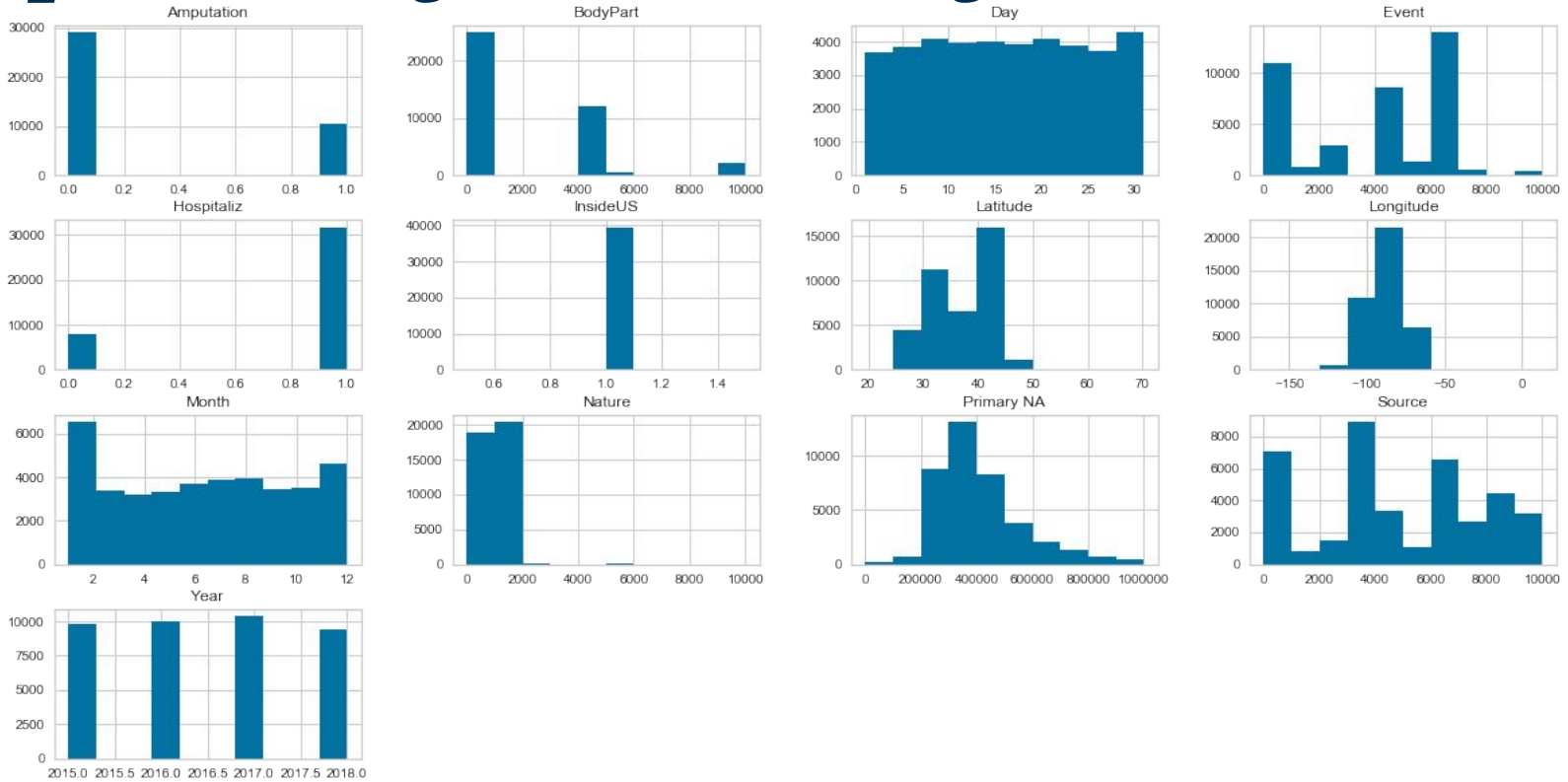
- Relied on two main data sources:
  - OSHA Severe Injury Report Data
  - U.S. Census TIGER data for zip codes, counties and data.

- Originally attempted to use zip-codes as a means to group data geographically.
  - Surprise! Not all incidents **in** the United States.

- Switched to counties using Latitude/Longitude and geospatial join to reassign county to each incident.
  - TIGER data crucial for this step

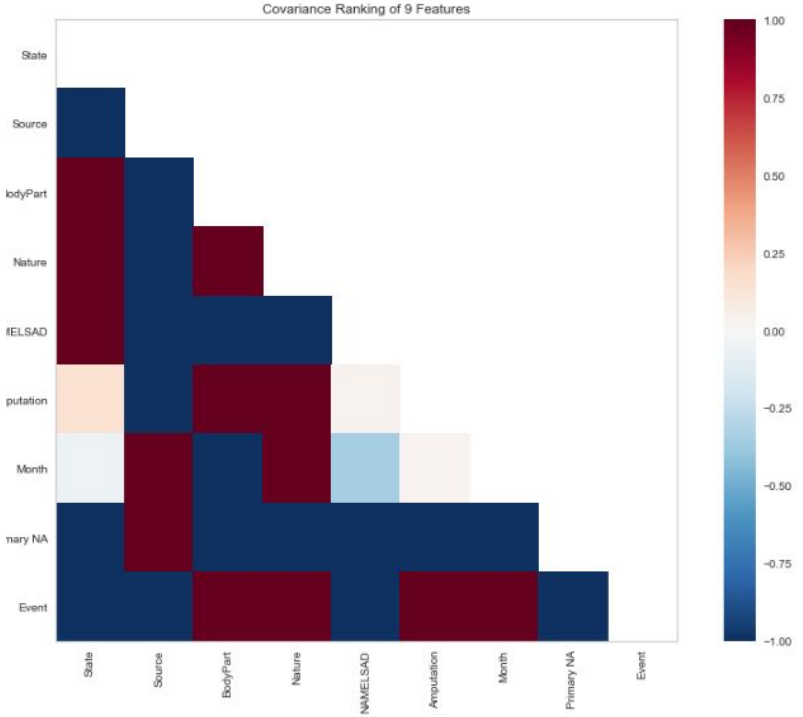


Exploratory Data Analysis



Exploratory Data Analysis

- Explored Yellowbrick visuals
  - Covariance graph on 9 selected features
  - Classification heatmap (not shown)
- Primary usage: Visualizing clustering algorithms



Feature Engineering

```
#Label Encode the two non-numerical value features
le = LabelEncoder()
data['State'] = le.fit_transform(data.State.values)
data['NAMELSAD_Codes'] = le.fit_transform(data.NAMELSAD.values)
```

Label encoding

Essential feature engineering step

Unsupervised Machine Learning

Clustering Analysis using Scikit Learn and YellowBrick

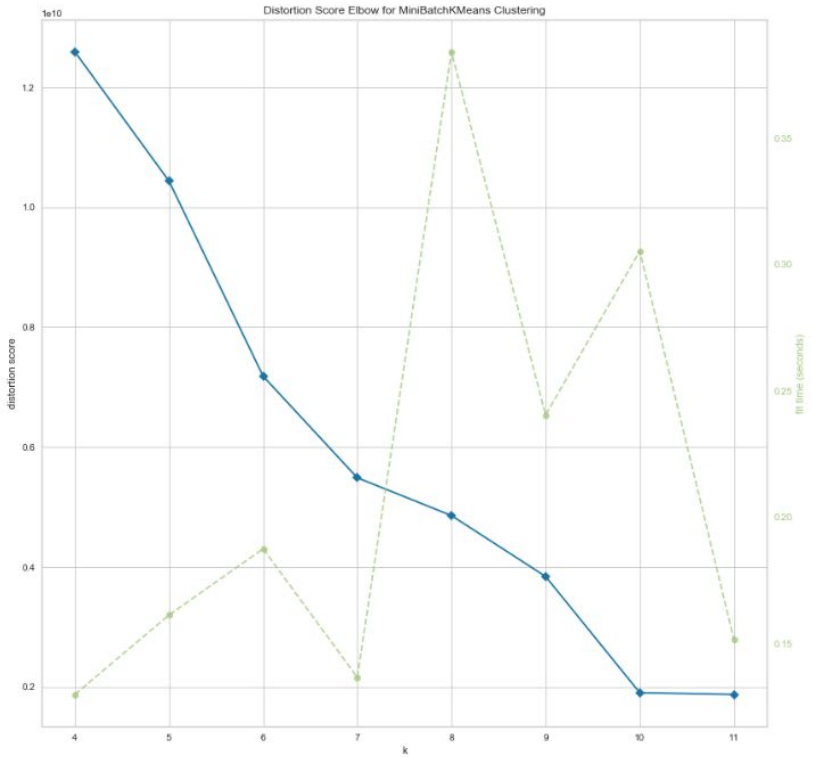
Tested alternative clustering models

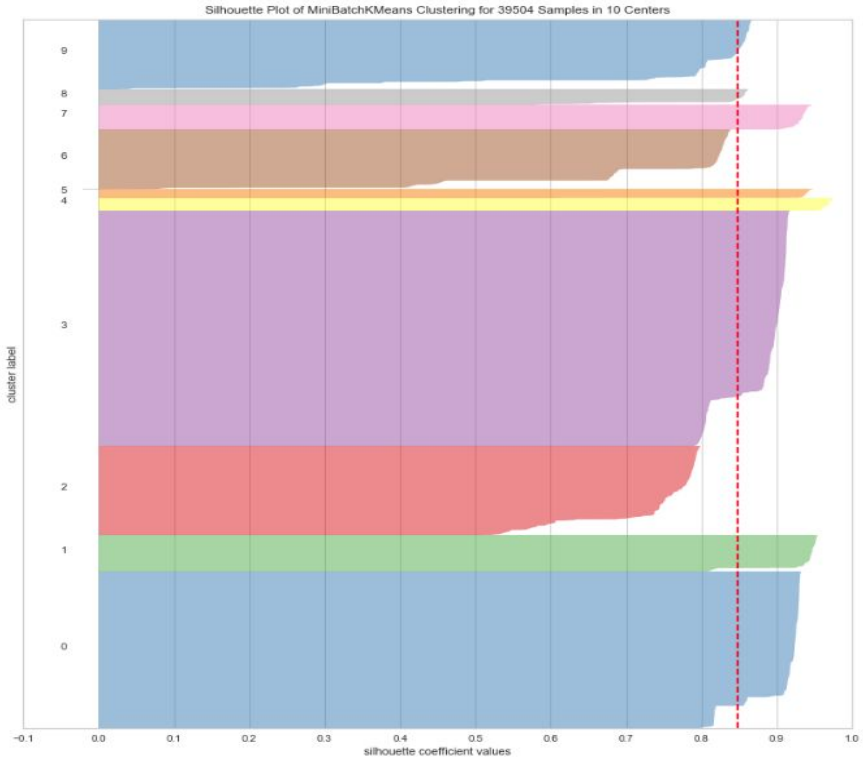
- K-means, Mini-Batch K-means, Agglomerative, DBScan, and Spectral
  - Spectral computationally ‘costly’, could not complete clustering.

Focused on K-means for primary analysis

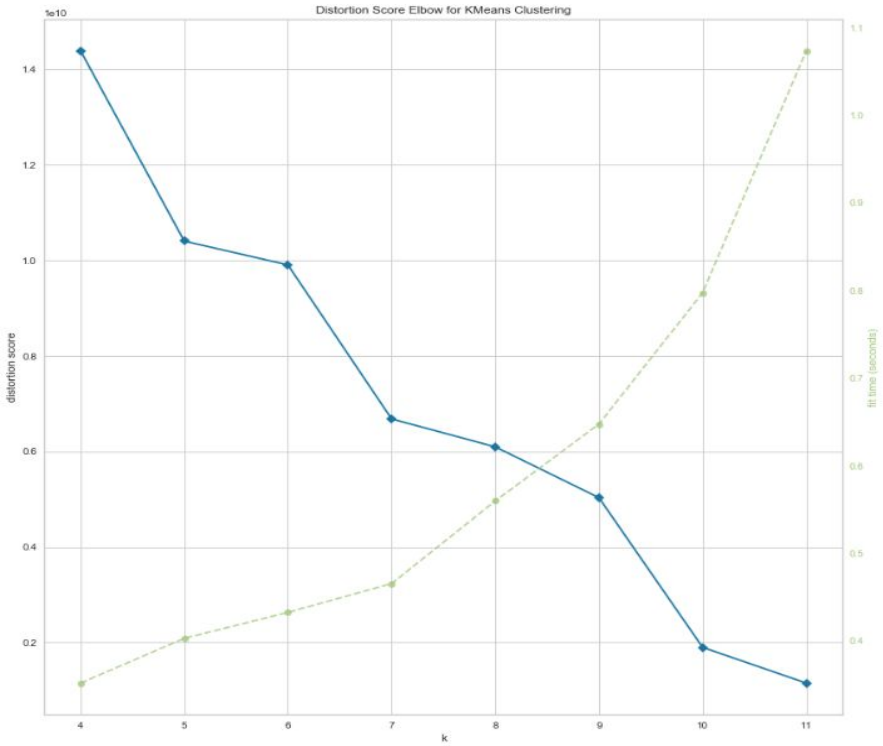
```
Method Name:KMeans, # of Clusters:10, Silhouette Score:0.8554840295513454
Method Name:MiniBatchKMeans, # of Clusters:10, Silhouette Score:0.7898094946314074
Method Name:AgglomerativeClustering, # of Clusters:10, Silhouette Score:0.8454500523138355
Method Name:AgglomerativeClustering, # of Clusters:10, Silhouette Score:0.8435510106779214
Method Name:AgglomerativeClustering, # of Clusters:10, Silhouette Score:0.7892567305002927
Method Name:DBSCAN, # of Clusters:10, Silhouette Score:0.7326845110830787
```

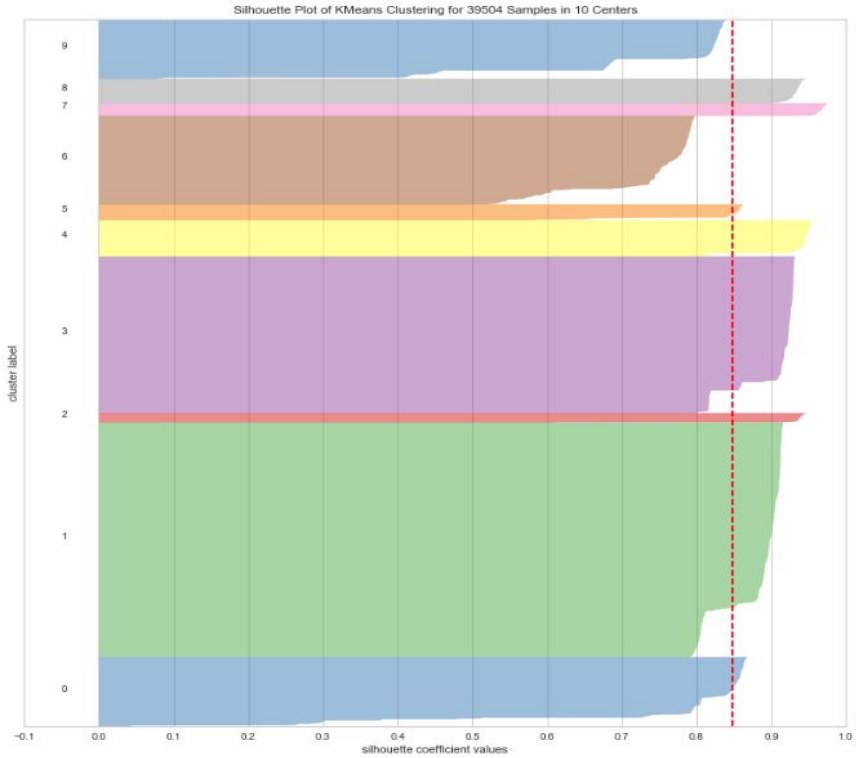
MiniBatch Analysis





K-Means Analysis



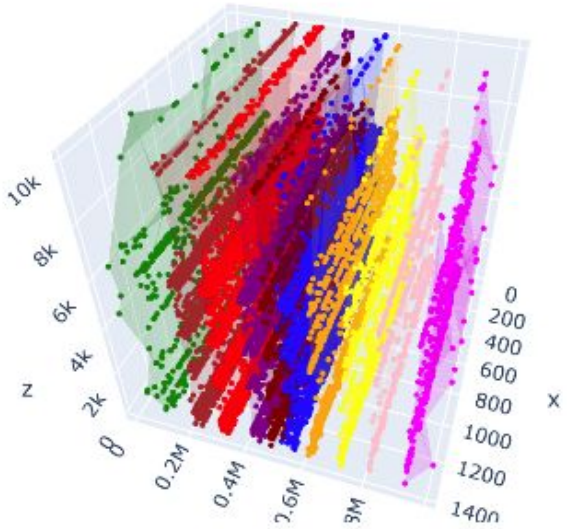


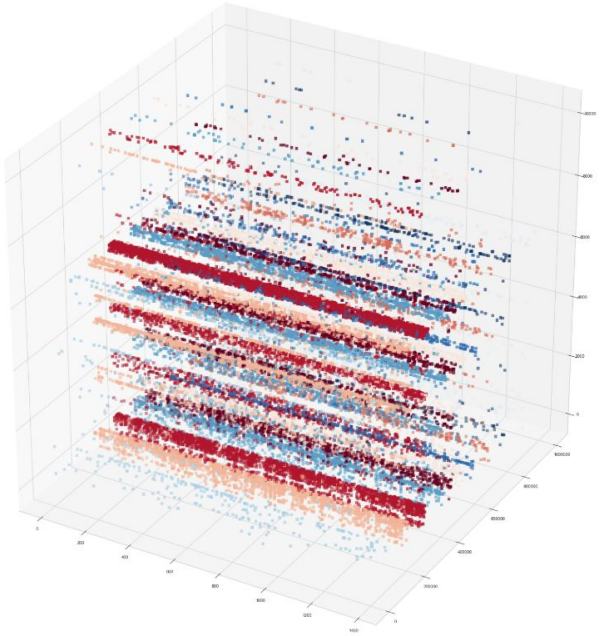
Results & Reflections

Results

Goal: Clustering by injury

Achieved: Clustering by *industry code* and *event* that caused the injury

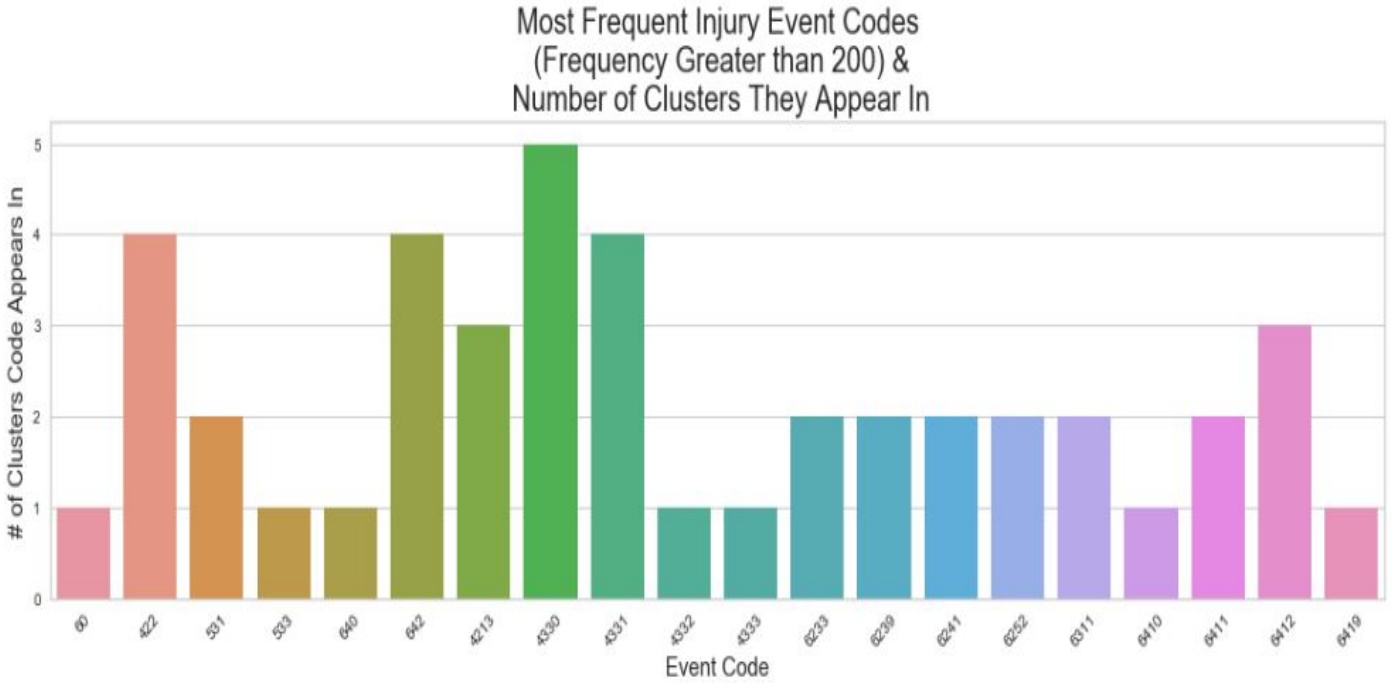




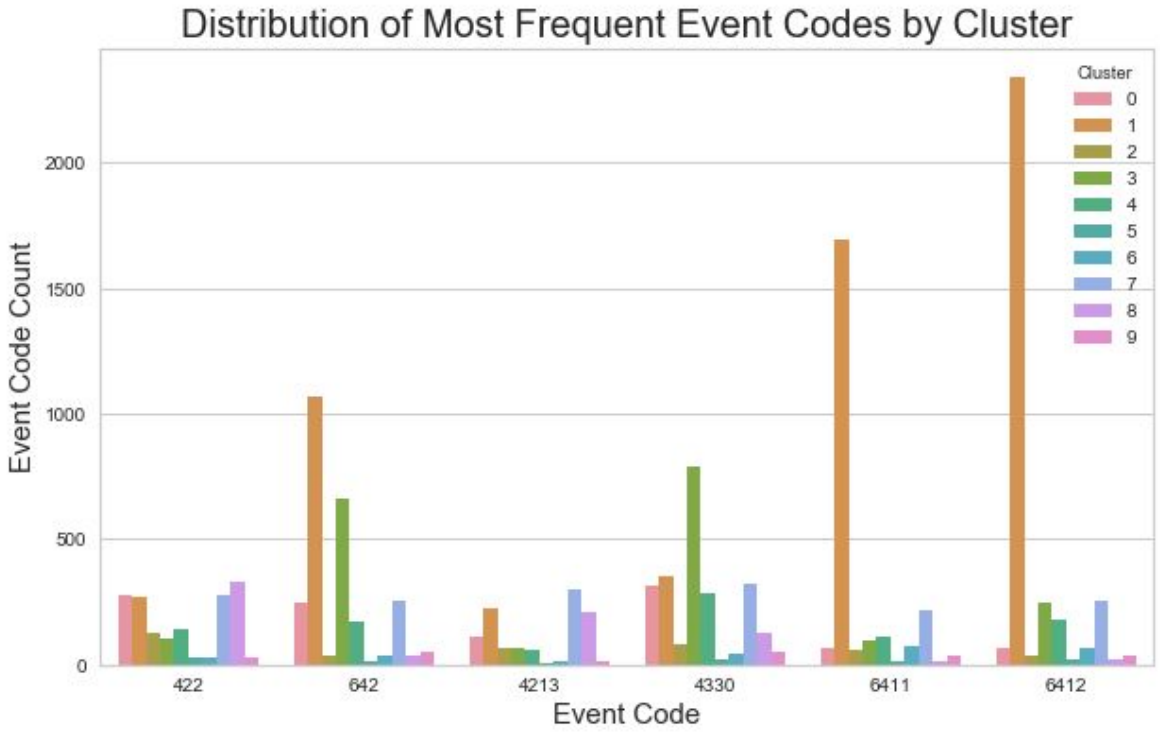
Results

Cluster #	***Primary Industry Represented***	***Cluster #***	***Primary Industry Represented***
0	Transportation & Warehousing	5	Public Administration
1	**Manufacturing**	6	Agriculture
2	Arts & Recreation	7	Retail Trade
3	Mining & Construction	8	Health Care
4	Waste & Remediation Services	9	Maintenance & Repair

Results


Most Frequent Injury Event Codes
(Frequency Greater than 200) &
Number of Clusters They Appear In
# of Clusters Code Appears In
0
1
2
3
4
5
60
422
531
533
640
642
4213
4330
4331
4332
4333
6233
6239
6241
6252
6311
6410
6411
6412
6419
Event Code

Results


Distribution of Most Frequent Event Codes by Cluster
Cluster
0
1
2
3
4
5
6
7
8
9
Event Code Count
0
500
1000
1500
2000
422
642
4213
4330
6411
6412
Event Code

Try, Try Again

A tale of pivoting…

- Our biggest lesson learned - thoroughly explore your intended dataset.
  - Revised data source 3 times due to nature of data.
  - Maintained injury focus - shift from Emergency Rooms to OSHA Severe Injuries

Next Steps

We believe we can reveal more information about the injuries by the following improvements:

- Add industry information to the data. Specifically type of industry, size, safety training data if available and location
  - For e.g. construction accounts for more than 20% of the fatalities
- We could improve the clustering by using Autoencoders (artificial neural networks) to reduce the dimensions and then apply k-means
- This method could be explored instead of using PCA especially to handle some of the text injury narratives

Would also consider a natural language processing (NLP) analysis

Thank you for your time!

We are happy to take any questions.

GEORGETOWN UNIVERSITY

Feedback, reactions, questions