IBM Unsupervised Machine Learning Assignment

By

Ahmed Shahriar Sakib

Introduction:

This is an end-to-end project on unsupervised machine learning which aims to get insights from a “local emergency incidents” Dataset and perform clustering techniques on several features.

Main Objective

The main objective of this project is to collect data, prepare it and perform exploratory data analysis which is followed by clustering and dimensionality techniques. The result from the analysis might be beneficial to a varieties of business stakeholders. For example – For real estate business agencies may take decision based on emergency occurrences and its frequency that which place is risky for housing, and which are not and take precaution properly. Or for local authority to avoid planting oil/gas filling stations fire prone locations. It will be also helpful for local government to estimate a proper budget to take preventive measures for local emergencies and other natural phenomenon.

Background

PulsePoint is a 911-connected mobile app that allows users to view and receive alerts on calls being responded to by fire departments and emergency medical services. The app's main feature, and where its name comes from, is that it sends alerts to users at the same time that dispatchers are sending the call to emergency crews. The goal is to increase the possibility that a victim in cardiac arrest will receive cardiopulmonary resuscitation (CPR) quickly. The app uses the current location of a user and will alert them if someone in their vicinity is in need of CPR. The app, which interfaces with the local government public safety answering point, will send notifications to users only if the victim is in a public place and only to users that are in the immediate vicinity of the emergency.

Pulsepoint logs of the incidents can be used to identify local pattern of emergencies which is helpful for local businesses as well as emergency agencies to stay alert and take precaution which, in the long term ensure social well-being of the people.

Dataset:

The dataset was collected via web scraping using selenium python library. The logs were collected from 2021-05-02 to 2021-11-16.

Source

- A private dataset containing records of incidents from https://www.pulsepoint.org

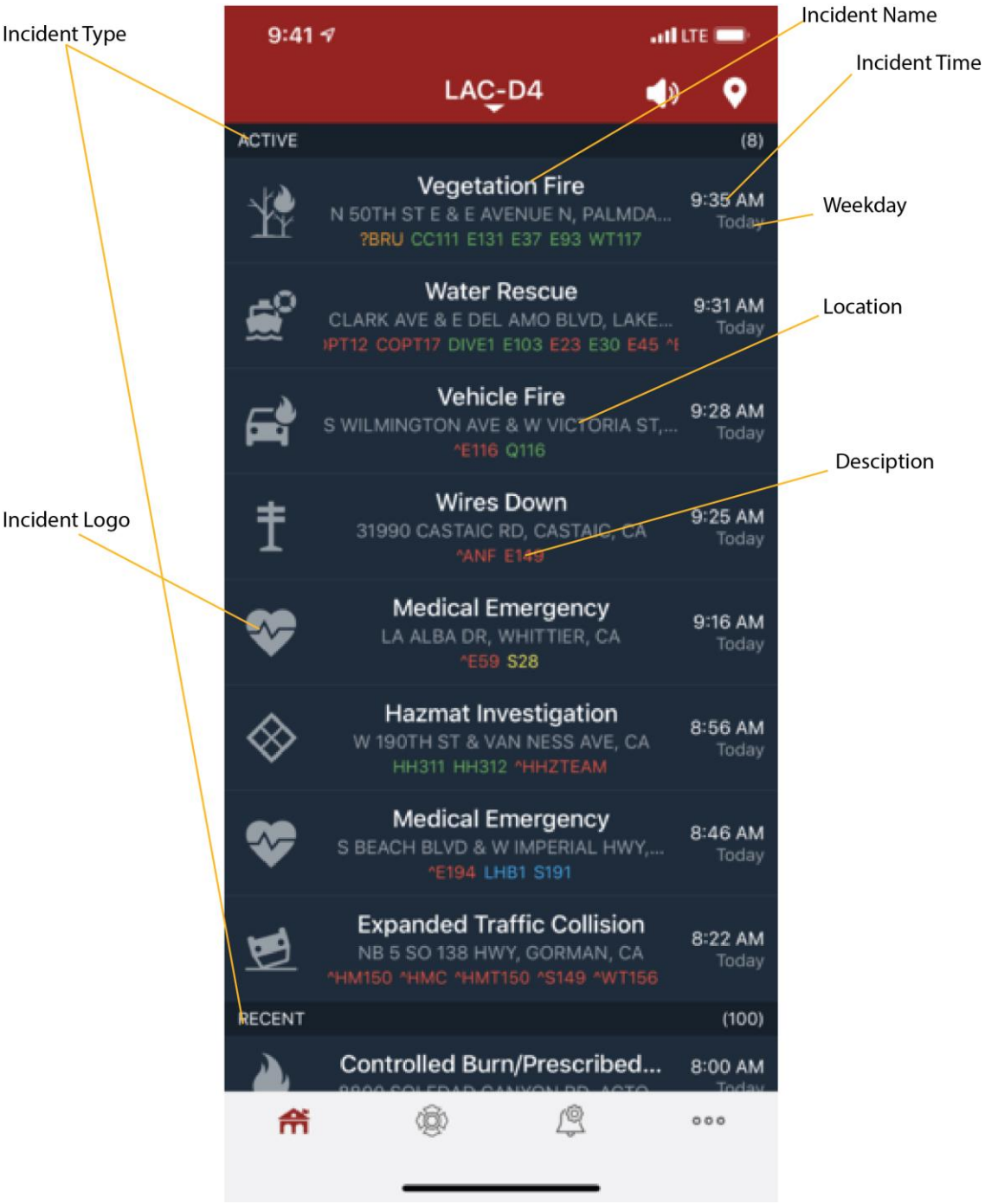


Figure: PulsePoint Mobile APP UI (visual inspection of the data)

Descriptive Analysis

Dataset dimension: **284306** rows × **13** columns

Data Types:

Data Type	Count
object	9
int64	1
float64	2
datetime64	1

Feature Summary

Columns	Description	Data Type
id	Contains record id	numeric, int
type	Incident type (recent or active)	object
title	Title of the incident	object
agency	Agency name	object
location	Location where the emergency took place	object
timestamp_time	Time when the emergency record was logged	object
date_of_incident	Date when emergency record was logged	datetime
description	Emergency code description	object
duration	Duration of the incident	object
Incident_logo	Logo of the incident	object
agency_logo	Logo of the agency	object
longitude	Longitude of the location	numeric, float
latitude	Latitude of the location	numeric, float

Object type Data description

	count	unique	top	frequency
type	284306	2	recent	220221
title	284306	89	Medical Emergency	187201
agency	284306	776	Montgomery County	5990
location	284306	184179	EUCLID AV, EUCLID, OH	109
timestamp_time	284306	1440	6:41 AM	351

	count	unique	top	frequency
description	271092	93698	E1	1058
duration	220221	682	15 m	5051

Numerical type Data description

	count	mean	std	min	25%	50%	75%	max
longitude	276334.0	-82.776968	45.091404	-178.140953	-115.118928	-86.528834	-76.881058	178.702778
latitude	276334.0	37.185910	11.957931	-90.000000	33.958496	38.869296	42.115496	74.964203

Data Exploration Plan

- Check for missing values, noise
- Discard irrelevant data
- Cleaning and preprocess dataset, impute missing values if necessary
- Extract features if applicable
- Location related data exploration
- Time series data exploration
- Perform clustering techniques, such as K-means

Data Cleaning and Feature Engineering

Missing Values

Features	Total	Percentage
duration	64085	22.54
description	13214	4.64
latitude	7972	2.80
longitude	7972	2.80

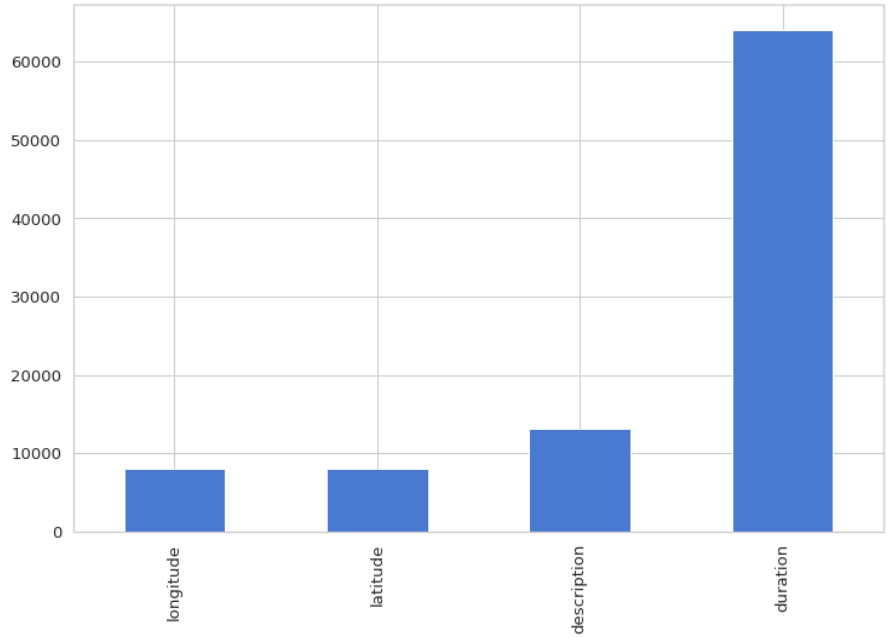


Figure: Missing Values Distribution

Data Cleaning & Feature Engineering

- Deleted unnecessary column ("**Id**", "**incident_logo**" and "**agency_logo**")
- Removed active incidents. Because active incidents are the noisy duplicated data of the "**recent type incident**" which was unable to remove during the data collection process. Thus, it does not contribute towards the analysis.
- The **location** column was divided up to maximum four columns – **address** , **address_2**, **city** and **state**
- Dropped garbage values from location column
- Mapped Canada provinces to their unique short form (**state** column)
- Dropped and mapped appropriate "**state"**
- Extracted duration total time from "**duration**" text and added new feature "duration in seconds"
- Added new feature "**time_of_the_day**" from "**timestamp_time**" feature
    - morning -> 5AM to 11:59AM
    - afternoon ->12PM to 4:59PM
    - evening -> 5PM to 8:59PM
    - night -> 9PM to 11:59PM
    - midnight -> 12AM to 4:59AM

Exploratory Data Analysis

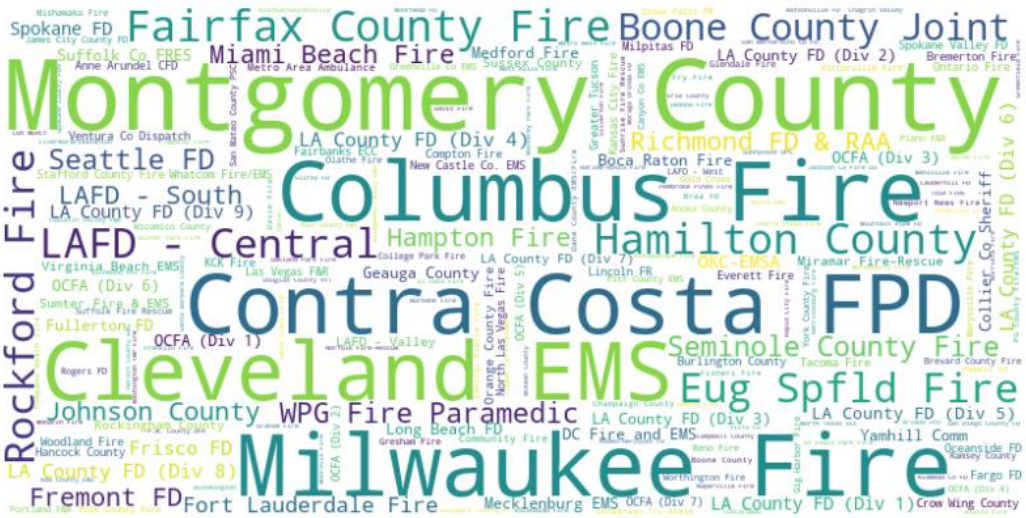
- After cleaning the dataset has **220219** entries and **87** type local emergencies (e.g., fire alarm, medical emergency, mutual aid etc.)

- There are **44** states (USA) / provinces (Canada) and **3389** cities in total
- Total **6479** locations were missing from geolocation data (geocode) which were discarded during geolocation data analysis
- Top 10 local emergencies –

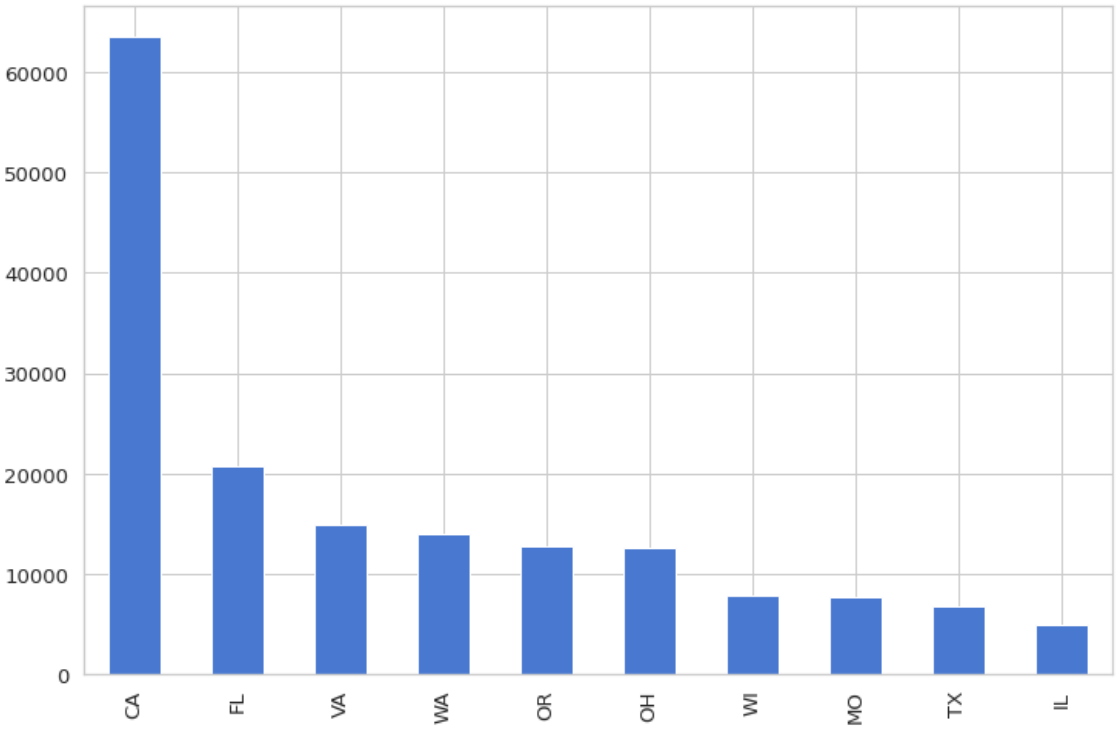
Incidents	Count
Medical Emergency	138891
Traffic Collision	17639
Fire Alarm	8543
Alarm	5909
Public Service	5605
Refuse/Garbage Fire	3988
Structure Fire	3728
Mutual Aid	2646
Fire	2500
Lift Assist	2363

- Word Cloud Incidents

- Word Cloud agencies – most frequent Montgomery County



- Number of incidents occurrences by states (highest “CA”- california)–



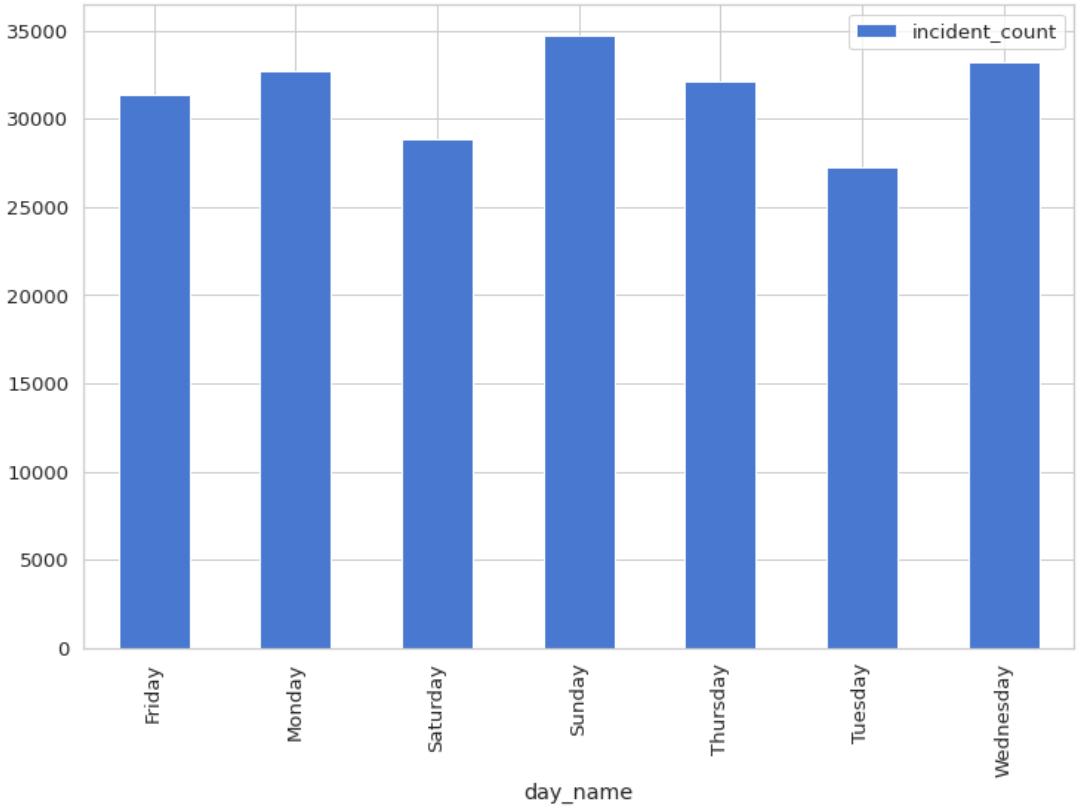
- Top agencies by incident engagement count -

Agency	**Incident Attended**
Montgomery County	5166
Milwaukee Fire	3813
Columbus Fire	3546
Contra Costa FPD	3485
Cleveland EMS	3375
Fairfax County Fire	2881
Hamilton County	2854
Eug Spfld Fire	2685
Rockford Fire	2614
LAFD - Central	2566
Boone County Joint	2423
Seminole County Fire	2334
Seattle FD	2292
WPG Fire Paramedic	2201
Richmond FD & RAA	2166

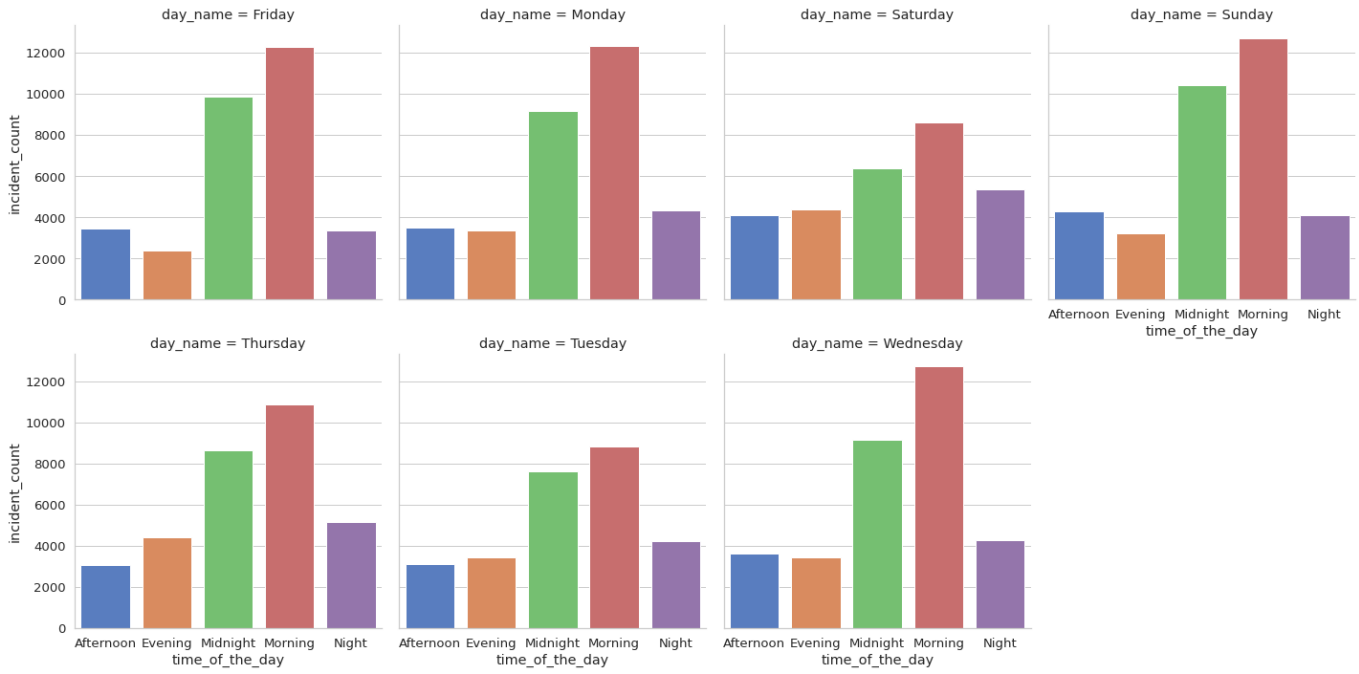
- Most incidents occurrences by dates –

Date	Number of Incidents
2021-11-13	3918
2021-11-07	3164
2021-08-20	2947
2021-06-22	2931
2021-06-18	2890
2021-06-16	2865
2021-11-04	2862
2021-08-23	2850
2021-10-12	2755
2021-09-16	2692

- Incidents by weekday (most in **Sunday**)-



- In terms of incidents occurred by week day :
    - Sunday > Wednesday > Friday > Monday > Thursday > Tuesday > Saturday



- Most of the incidents occurs during Midnight or in the morning. Probably some of the incidents were already started at nighttime and were logged later in the morning.

Top ten emergencies during 'Midnight' or 'Morning' -

Midnight:

- Medical Emergency
- Traffic Collision
- Fire Alarm
- Alarm
- Public Service
- Structure Fire
- Refuse/Garbage Fire
- Mutual Aid
- Residential Fire
- Expanded Traffic Collision

Morning:

- Medical Emergency

- Traffic Collision
- Fire Alarm
- Public Service
- Refuse/Garbage Fire
- Structure Fire
- Fire
- Residential Fire
- Mutual Aid
- Lift Assist

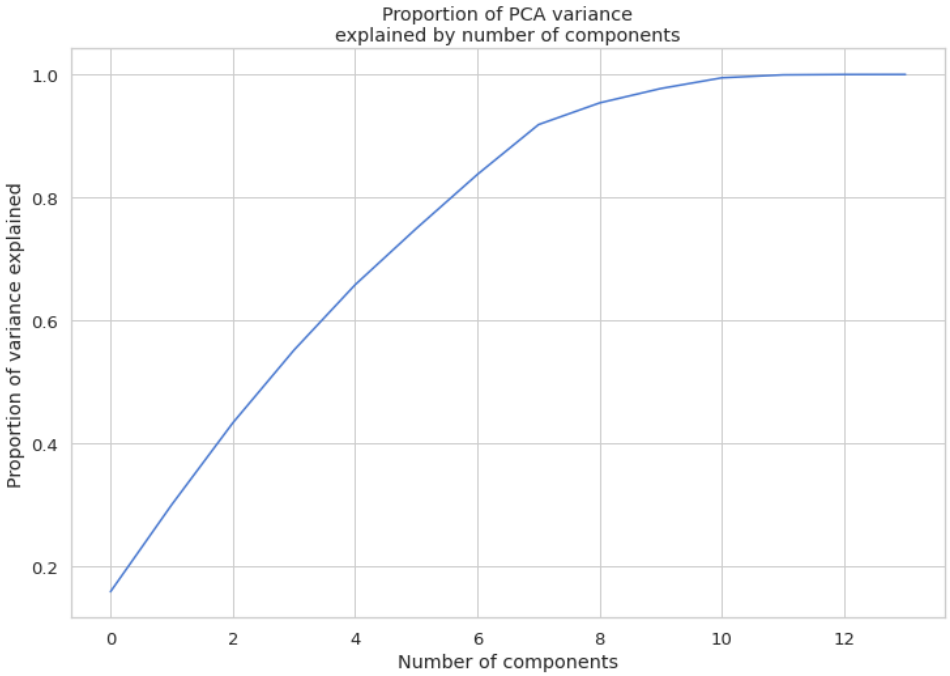
Dimensionality Reduction

PCA

To prepare the dataset for unsupervised analysis, features such as - "type", "timestamp_time", "date_of_incident" , "business", "address_2" and "duration" was removed; The data set was left with 16 features in total.

Remarks –

- Type only contains 'recent' which is an unnecessary feature
- "timestamp_time" was replaced with "time_of_the_day" feature
- "date_of_incident" was replaced with "week_day", "day_name" and "month_name"
- "business" and "address_2" has lots of null values, hence those features were removed
- "duration" was converted to numerical value and replaced with "duration_in_seconds"

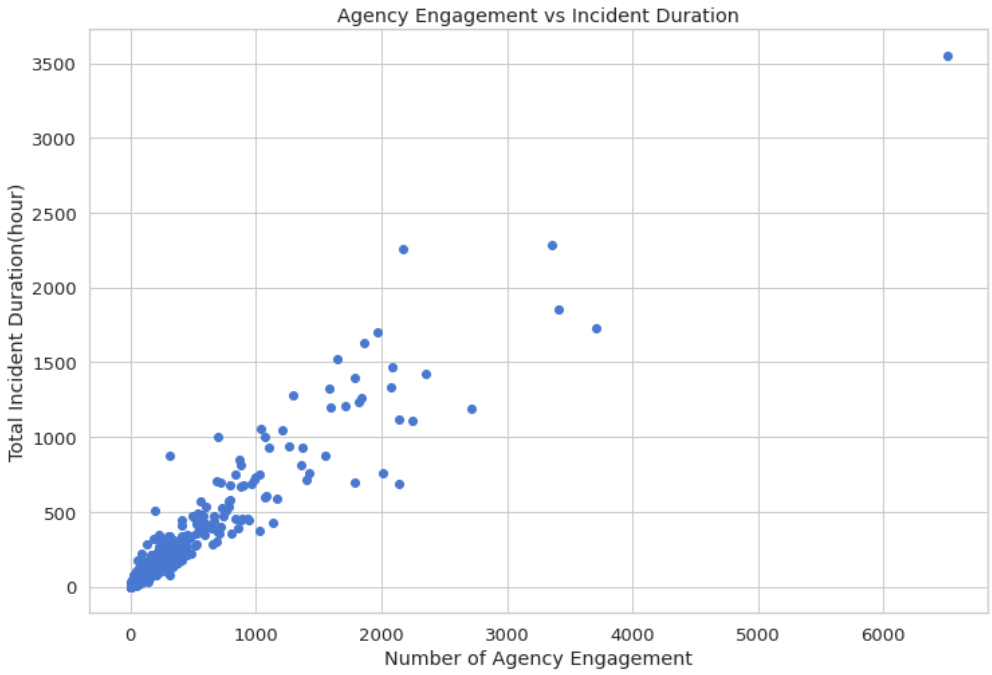


- From the above graph it's certain that 7 components are needed in order to explain ~90% of the variance in the data

Clustering

Agency Engagement Vs Incident Duration

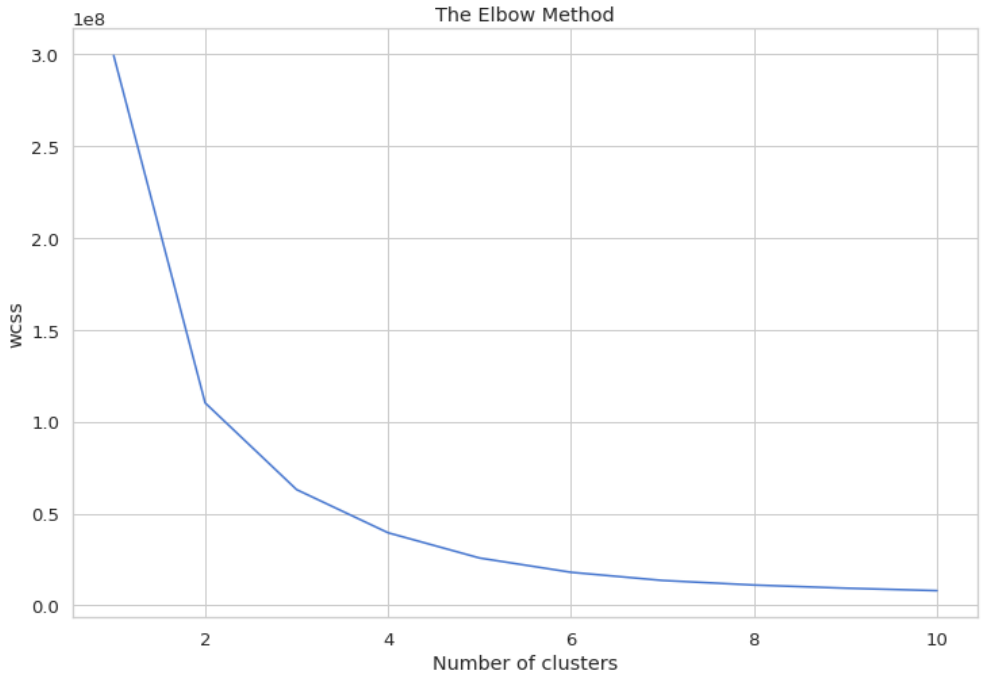
The **number of agency engagement** count and **duration(hour)** has a positive linear relationship. Higher **Duration of Incidents** indicates **more agency engagement** in a city

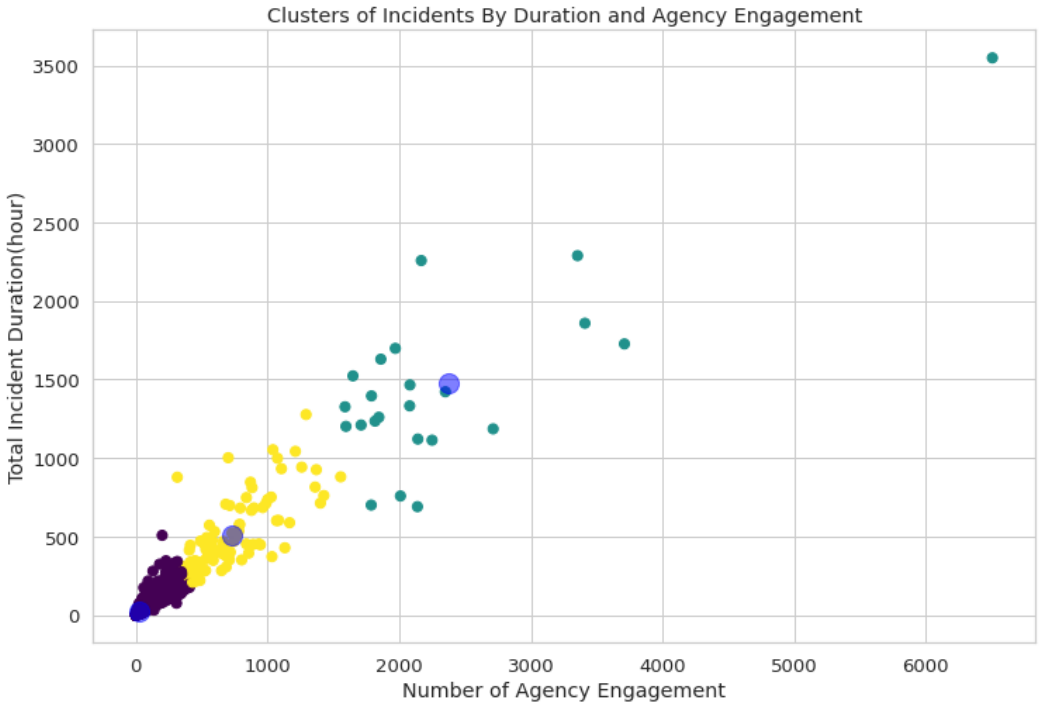


K-means

With “kmeans++” initialization the objective of this clustering is to create groups based on number of agency engagements and total incident duration (hour)

- The best number of K lies between 2 and 4





The k-mean clustering algorithm clusters the **cities** based on *duration of incidents* and *number of agencies* into **three groups**. Small duration indicates having less number of agency engagement and vice-versa.

Findings

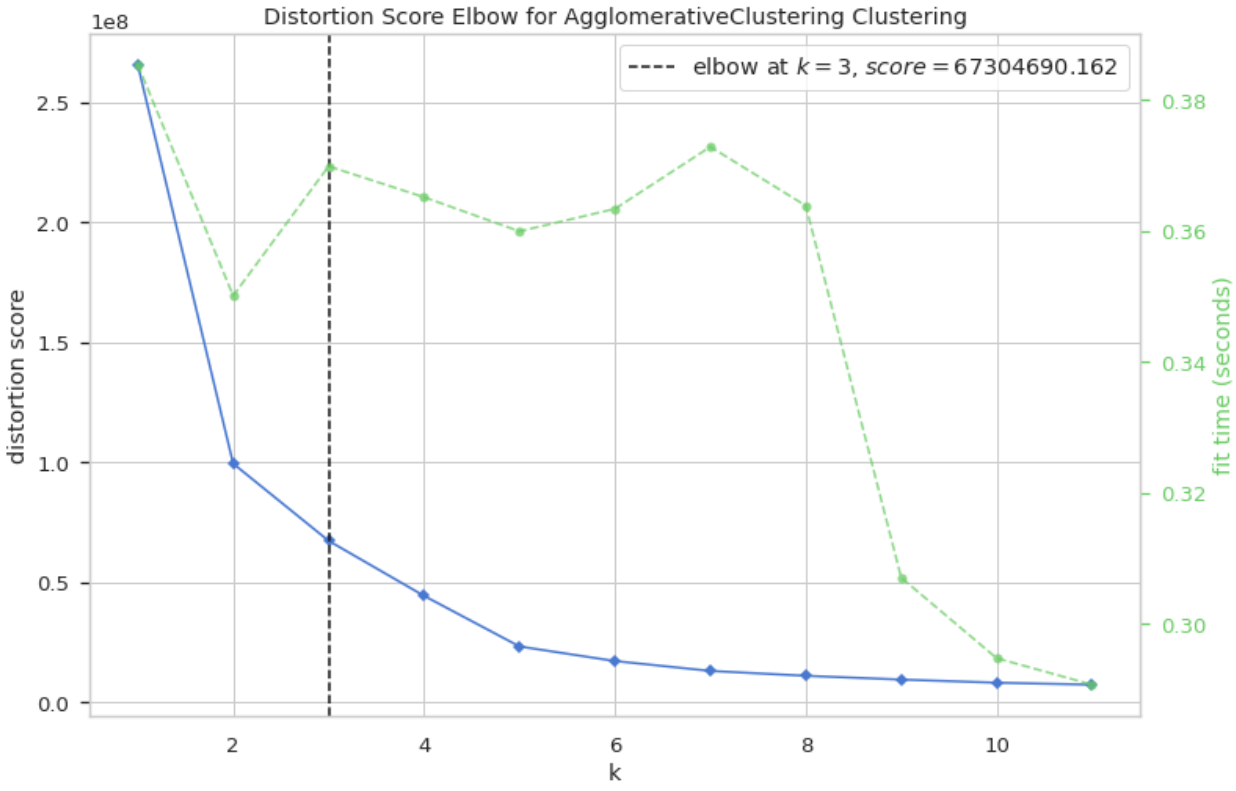
Group 1 : Cities with very low number of incidents duration and agency engagements

Group 2 : Cities with comparatively higher number of incidents duration and agency engagements

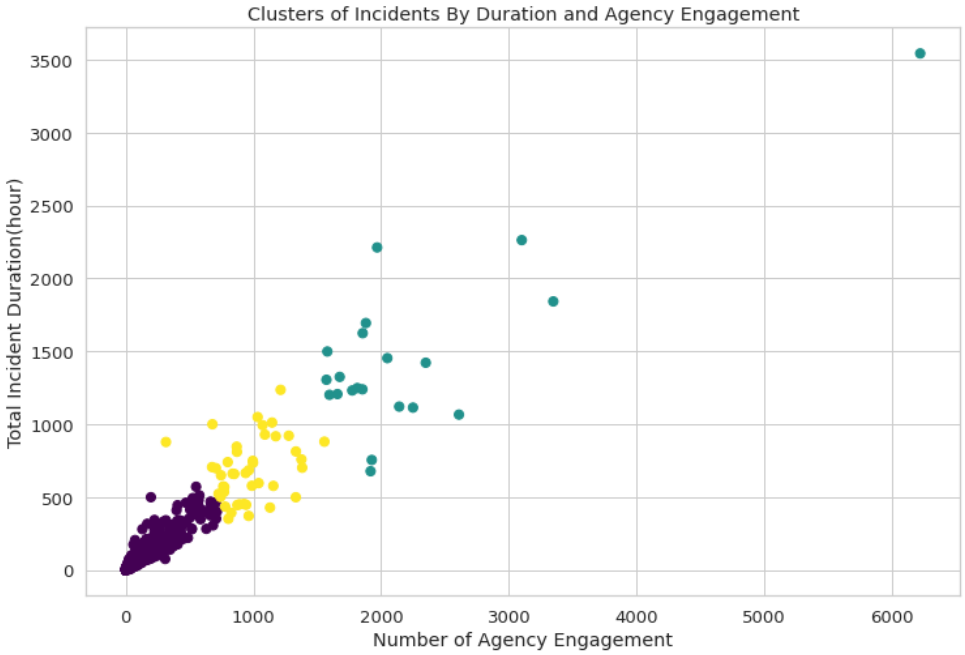
Group 2 : Cities with highest number of incidents duration and agency engagements

Agglomerative Clustering

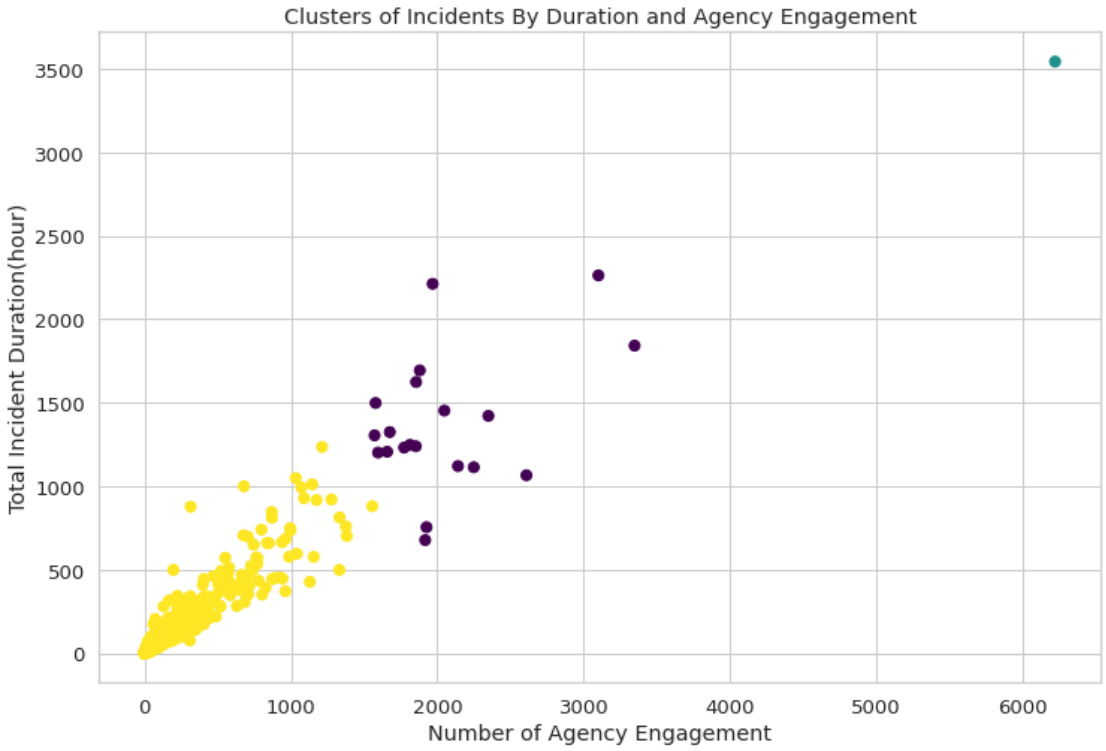
With “KElbowVisualizer” from yellowbrick library it’s found that optimum K value is 3 (it implements the “elbow” method)



- Ward linkage (clusters = 3)



- Complete Linkage (clusters =3)



Results

From the above clustering techniques, it is clear that "complete" linkage is not suitable for Agglomerative clustering (cluster parameter was given 3 but it returned 2 clusters). On the other hand, k-means and "Ward" Agglomerative provided a better clustering result. But the density of the cities is high when the value of number of agency engagement and total incident duration is low. K-means focused on clustering lower dense cities with unequal parameter distribution –

- **Cluster 0**: Total incident duration = ~400, number of engagements = ~500
- **Cluster 1**: Total incident duration = 400 to ~1200, number of engagements = 500 to ~1500
- **Cluster 2**: Total incident duration = 1200 to max, number of engagements = 1500 to max

The range of cluster 1 is bigger than cluster 0 in k-means whereas Ward Agglomerative did almost an equally distributed clustering for cluster 0 and 1. If the range of the parameters (engagements of duration) is important based on other factor, for example – budget allocation with respect to engagements or business decision/future planning based on duration of emergencies, then depending on the priority both clusters would be acceptable.

Key Insights

- Most of the incidents occurred in California
- Most incidents happened during midnight and in the morning throughout the week
- The highest number of incidents happened on Sunday
- The incidents' number got increased after Covid-19 lockdown
- Medical emergency was the highest occurring incident which was followed by traffic collision and fire alarm
- Montgomery County, Milwaukee Fire and Columbus Fire were the top active agencies during the five months period

Conclusion & Future Work

- More clustering techniques such as DBSCAN, t-SNE will be explored
- Different features will be considered for clustering
- Better result could be achieved by performing dimensionality reduction for clustering or hyperparameter tuning.