Outlier Detection

Cong Liu
New York University
New York, New York
cl4055@nyu.edu

Nan Su
New York University
Long Island City, New York
ns3783@nyu.edu

Hanlu Zhang
New York University
Jersy City, New Jersy
hz1625@nyu.edu

ACM Reference Format:
Cong Liu, Nan Su, and Hanlu Zhang. 1997. Outlier Detection. In *Proceedings of ACM Woodstock conference (WOODSTOCK'97).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Data plays a very important role as it collects facts in different formats(words, numbers, text, etc..), which have been translated into forms that can be processed by computers and thus can help get insights and make decisions. Under such circumstance, data quality, which refers to the degree to which a set of characteristics of data fulfills requirements, is the key for performing further analysis of a data set. However, in most cases, a data set has different kinds of data quality issues, and one of the most common issues is the outliers. The definition of outliers of Grubbs [4] and quoted in Barnett & Lewis [2]: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. A further outlier definition from Barnett & Lewis [2] is: An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.
There are many reasons that may cause outliers on a data set, such as:

- Data entry errors, which are caused by human errors.
- Measurement errors, which are caused by instrument errors.
- Experimental errors, which are caused by data extraction or experiment planning/executing errors.
- Data processing errors which are caused by data manipulation or data set unintended mutations.
- Sampling errors which are caused by extracting or mixing data from wrong or various sources

In addition, there are always entries with null values on a data set that may affect the analysis as well. Discovering these entries is one of the main job for data cleaning, and it is also widely used for different applications, such as fraud detection, which detects fraudulent applications for credit cards, state benefits or detects fraudulent usage of credit cards or mobile phones, loan application processing, which detects fraudulent applications or potentially problematical customers, intrusion detection, which detects unauthorized access in computer networks, etc[5].

If we can use efficient algorithms to find out outliers and null values at the very beginning of data processing, it can save a lot of time and give us more accurate results from data analysis.

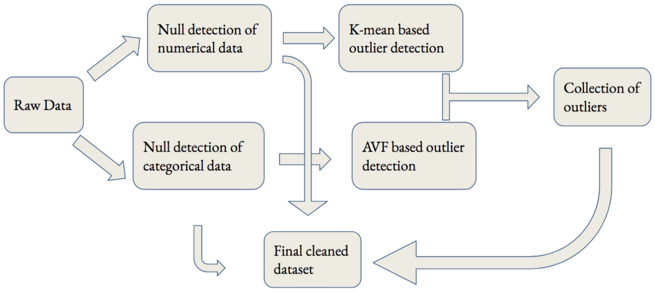
2 PROBLEM FORMULATION

Our goal for the project is to discover outliers from any given large data set. The problem we are interested in is how to implement some algorithms proposed in previous work and design or improve algorithms that can return more accurate results and achieve our goal in a more efficient way. Since there are different metrics to define outliers under different contexts, we focus on methods both for numerical variables and categorical variables.

3 METHODS, ARCHITECTURE AND DESIGN

In this section we want to introduce some common and implements which we will use as baseline algorithms. Detecting outliers in the data is not as explicit as it sounds like. It is hard to define outlier for a single attribute. For instance, when it comes to the traffic sensor data, a speed of 60 mph may not be anomaly if the vehicle is on expressway but it is definitely strange if the car is running in a congested city like New York. In this case, volumn and speed should be considered at the same time.

The whole procedure of detecting null values and outliers in our project is as following:


Raw Data
Null detection of numerical data
Null detection of categorical data
K-mean based outlier detection
AVF based outlier detection
Collection of outliers
Final cleaned dataset

3.1 Categorical data

Attribute value frequency(AVF) method There are various algorithms dealing with categorical data such as The Greedy Algorithm. However it is linear with respect to data size and it needs k-scans each time. The other models also exist which are based on frequent item set mining(FIM) need to create a large space to store item sets, and then search for these sets in each and every data point .These techniques can become very slow when we select low threshold value to find frequent item sets from dataset.

Intuitively, outliers are those points which are infrequent in the dataset. The infrequentness of an attribute value can be measured by computing the number of times this value is assumed by the corresponding attribute in the dataset.

Another simpler and faster approach to detect outliers that minimizes the scans over the data and does not need to create more space and more search for combinations of attribute values or item sets is Attribute Value Frequency (AVF) algorithm.

WOODSTOCK'97, July 1997, El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

Assuming that a dataset contains n data points, $x_i, i = 1, fi, n$. If each data point has m attributes, $x_i = [x_{i1}, ..., x_{ij}, ..., x_{im}]$, where x_{ij} is the value of the j-th attribute of x_i.

An outlier point x_i is defined based on the AVF Score below:

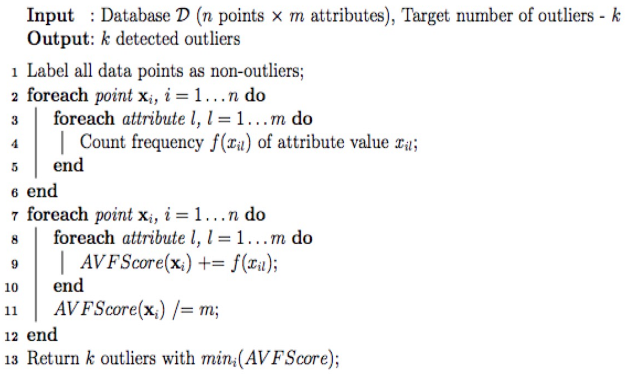
$$AVFScore(x_i) = F_i = \frac{1}{m} \sum_{j=1}^{m} f(x_{ij})$$

where $f(x_{ij})$ is the number of times the j-th attribute value of x_i appears in the dataset. Since we essentially have a sum of m positive numbers, the AVF score is minimized when each of the summation terms are individually minimized. Thus, the AVF score will be minimum for the ideal outlier.

Once we calculate the AVF score of all the points, we can define the k points with the smallest AVF scores as the k outliers. The complexity of AVF is $O(n \times m)$ compared to Greedy's complexity, $O(k \times n \times m)$ since AVF detects outlier after only one scan of the dataset rather than k scans, which is needed by Greedy Algorithm.

Specifically, AVF exhibits the following advantages:

- AVF needs one dataset scan to detect the desired outliers. In addition, the number of data scans needed by AVF does not rely on the input number of target outliers k.
- A computational complexity of $O(n \times m)$, where n is the number of points in the data and m is the data dimensionality. It scales linearly with both n and m.
- The performance of AVF is independent on additional user-entered parameters, such as minimum support.
- AVF runs significantly faster than the existing representative techniques, while maintaining comparable detection accuracy.

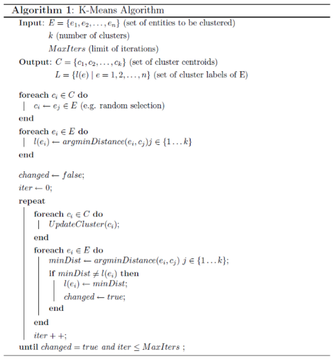


```
Input  : Database $\mathcal{D}$ ($n$ points × $m$ attributes), Target number of outliers - $k$
Output: $k$ detected outliers
Label all data points as non-outliers;
foreach point $\mathbf{x}_i$, $i = 1 \ldots n$ do
    foreach attribute $l$, $l = 1 \ldots m$ do
        Count frequency $f(x_{il})$ of attribute value $x_{il}$;
    end
end
foreach point $\mathbf{x}_i$, $i = 1 \ldots n$ do
    foreach attribute $l$, $l = 1 \ldots m$ do
        $AVFScore(\mathbf{x}_i)$ += $f(x_{il})$;
    end
    $AVFScore(\mathbf{x}_i)$ /= $m$;
end
Return $k$ outliers with $min_i(AVFScore)$;
```

Figure 1: AVF Algorithm

3.2 Numerical data

K-means method Data clustering is used to identify homogeneous groups or clusters from a set of objects. In other words, data clustering aims to divide a set of objects into groups or clusters such that objects in the same cluster are more similar to each other than to objects from other clusters[1]. As an unsupervised learning process, data clustering is often used as a preliminary step for data analytics. Among the many clustering algorithms have been developed in the past years, the k-means algorithm is one of the oldest and commonly used clustering algorithms. In the k-means based outlier detection technique, the data are partitioned into k groups by assigning them to the closest cluster centers. Once assigned we can compute the distance or dissimilarity between each object and its cluster center, and pick those with largest distances as outliers. Although being used widely, the k-means algorithm has some drawbacks[3]. One of the drawbacks is that it is extremely sensitive to outliers, and such outliers may have a disproportionate impact on the final cluster configuration.



```
Algorithm 1: K-Means Algorithm
Input: $E = \{e_1, e_2, \ldots, e_n\}$ (set of entities to be clustered)
       $k$ (number of clusters)
       $MaxIters$ (limit of iterations)
Output: $C = \{c_1, c_2, \ldots, c_k\}$ (set of cluster centroids)
        $L = \{l(e) \mid e = 1, 2, \ldots, n\}$ (set of cluster labels of E)

foreach $c_i \in C$ do
    $c_i \leftarrow e_j \in E$ (e.g. random selection)
end
foreach $e_i \in E$ do
    $l(e_i) \leftarrow argminDistance(e_i, c_j) j \in \{1 \ldots k\}$
end

$changed \leftarrow false$;
$iter \leftarrow 0$;
repeat
    foreach $c_i \in C$ do
        $UpdateCluster(c_i)$;
    end
    foreach $e_i \in E$ do
        $minDist \leftarrow argminDistance(e_i, c_j)\ j \in \{1 \ldots k\}$;
        if $minDist \neq l(e_i)$ then
            $l(e_i) \leftarrow minDist$;
            $changed \leftarrow true$;
        end
    end
    $iter++$;
until $changed = true\ and\ iter \leq MaxIters$ ;
```

Figure 2: K-means Algorithm

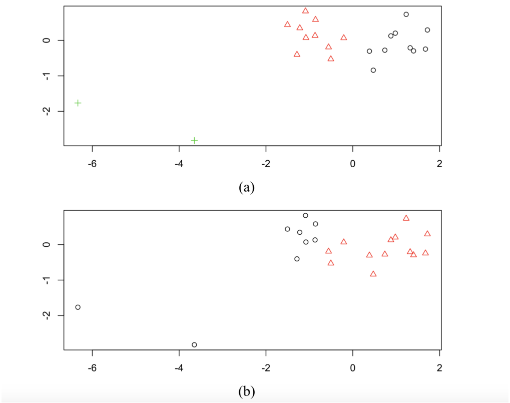


Figure 3: An illustration showing that the k-means algorithm is sensitive to outliers. (a) A data set with two clusters and two outliers. The two clusters are plotted by triangles and circles, respectively. The two outliers are denoted by plus signs. (b) Two clusters found by the k-means algorithm. The two found clusters are plotted by triangles and circles, respectively.

4 RESULTS

4.1 Standard AVF and Kmeans

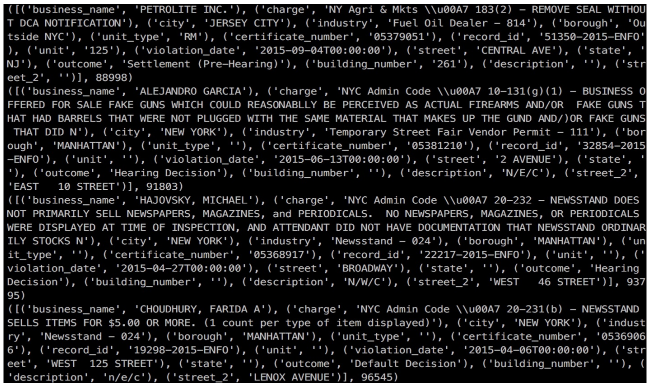
We implemented these algorithms using PySpark on 50 datasets from NYC open data. We ran all our experiments on Dumbo - a

Hadoop Cluster provided by High Performance Computing(HPC) at NYU.

Both algorithms were applied on all datasets. Given an arbitrary dataset, we first classify each attributes as categorical or numerical. Then the algorithms are implemented on corresponding attributes(clustering on numerical data and AVF on categorical data).

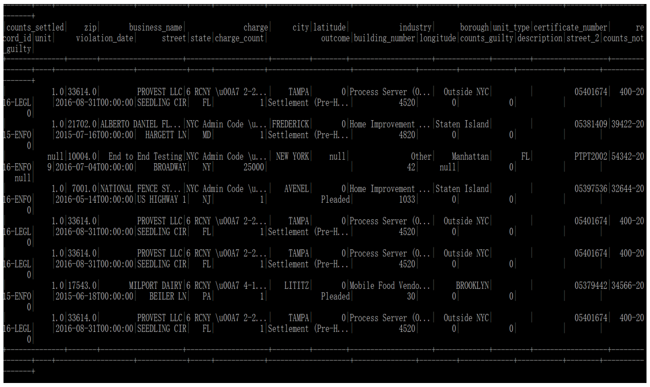
The 50 datasets have various structures and sizes. The largest dataset has 63867610 rows and 3 attributes while the smallest dataset has only 3 rows for which there is no need to detect outliers.

Taking one dataset as example, this dataset contains both categorical and numeric data which is an ideal target to demonstrate our outliers detection method. The data includes counts settled, zip, business name, charge, city, latitude, industry, borough, unit type, certificate number, record id, unit, violation date, street, state, charge count, outcome, building number, longitude, counts guilty, description, street2, counts not guilty as attributes. After applying the AVF algorithm with *k*(the number of outliers) set to 10, we get result as following(shown partially):



```
([('business_name', 'PETROLITE INC.'), ('charge', 'NY Agri & Mkts \\u00A7 183(2) - REMOVE SEAL WITHOU
T DCA NOTIFICATION'), ('city', 'JERSEY CITY'), ('industry', 'Fuel Oil Dealer - 814'), ('borough', 'Ou
tside NYC'), ('unit_type', 'RM'), ('certificate_number', '05379051'), ('record_id', '51350-2015-ENFO'
), ('unit', '125'), ('violation_date', '2015-09-04T00:00:00'), ('street', 'CENTRAL AVE'), ('state', '
NJ'), ('outcome', 'Settlement (Pre-Hearing)'), ('building_number', '261'), ('description', ''), ('str
eet_2', '')], 88998)
([('business_name', 'ALEJANDRO GARCIA'), ('charge', 'NYC Admin Code \\u00A7 10-131(g)(1) - BUSINESS O
FFERED FOR SALE FAKE GUNS WHICH COULD REASONABLLY BE PERCEIVED AS ACTUAL FIREARMS AND/OR  FAKE GUNS T
HAT HAD BARRELS THAT WERE NOT PLUGGED WITH THE SAME MATERIAL THAT MAKES UP THE GUND AND/)OR FAKE GUNS
 THAT DID N'), ('city', 'NEW YORK'), ('industry', 'Temporary Street Fair Vendor Permit - 111'), ('bor
ough', 'MANHATTAN'), ('unit_type', ''), ('certificate_number', '05381210'), ('record_id', '32854-2015
-ENFO'), ('unit', ''), ('violation_date', '2015-06-13T00:00:00'), ('street', '2 AVENUE'), ('state', '
'), ('outcome', 'Hearing Decision'), ('building_number', ''), ('description', 'N/E/C'), ('street_2',
'EAST   10 STREET')], 91803)
([('business_name', 'HAJOVSKY, MICHAEL'), ('charge', 'NYC Admin Code \\u00A7 20-232 - NEWSSTAND DOES
NOT PRIMARILY SELL NEWSPAPERS, MAGAZINES, and PERIODICALS.  NO NEWSPAPERS, MAGAZINES, OR PERIODICALS
WERE DISPLAYED AT TIME OF INSPECTION, AND ATTENDANT DID NOT HAVE DOCUMENTATION THAT NEWSSTAND ORDINAR
ILY STOCKS N'), ('city', 'NEW YORK'), ('industry', 'Newsstand - 024'), ('borough', 'MANHATTAN'), ('un
it_type', ''), ('certificate_number', '05368917'), ('record_id', '22217-2015-ENFO'), ('unit', ''), ('
violation_date', '2015-04-27T00:00:00'), ('street', 'BROADWAY'), ('state', ''), ('outcome', 'Hearing
Decision'), ('building_number', ''), ('description', 'N/W/C'), ('street_2', 'WEST   46 STREET')], 937
95)
([('business_name', 'CHOUDHURY, FARIDA A'), ('charge', 'NYC Admin Code \\u00A7 20-231(b) - NEWSSTAND
SELLS ITEMS FOR $5.00 OR MORE. (1 count per type of item displayed)'), ('city', 'NEW YORK'), ('indust
ry', 'Newsstand - 024'), ('borough', 'MANHATTAN'), ('unit_type', ''), ('certificate_number', '0536906
6'), ('record_id', '19298-2015-ENFO'), ('unit', ''), ('violation_date', '2015-04-06T00:00:00'), ('str
eet', 'WEST  125 STREET'), ('state', ''), ('outcome', 'Default Decision'), ('building_number', ''), (
'description', 'n/e/c'), ('street_2', 'LENOX AVENUE')], 96545)
```

Figure 4: AVF Result



```
+--------------+-------+------------------+-----------------+---------+--------+-----------------+-----------+---------+------------------+--------
-------+
|counts_settled|    zip|     business_name|           charge|     city|latitude|         industry|    borough|unit_type|certificate_number|      re
cord_id|unit|    violation_date|     street|state|charge_count|         outcome|building_number|longitude|counts_guilty|description|street_2|counts_not
_guilty|
+--------------+-------+------------------+-----------------+---------+--------+-----------------+-----------+---------+------------------+--------
-------+----+------------------+-----------+-----+------------+----------------+---------------+---------+-------------+-----------+--------+----------
-------+
|           1.0|33614.0|       PROVEST LLC|6 RCNY \u00A7 2-2...|    TAMPA|       0|Process Server (O...|  Outside NYC|         |          05401674|  400-20
16-LEGL|    |2016-08-31T00:00:00|SEEDLING CIR|   FL|           1|Settlement (Pre-H...|           4520|        0|            0|           |        |
      0|
|           1.0|21702.0|ALBERTO DANIEL FL...|NYC Admin Code \u...|FREDERICK|       0|Home Improvement ...|Staten Island|         |          05381409|39422-20
15-ENFO|    |2015-07-16T00:00:00|  HARGETT LN|   MD|           1|Settlement (Pre-H...|           4820|        0|            0|           |        |
      0|
|          null|10004.0|  End to End Testing|NYC Admin Code \u...| NEW YORK|    null|             Other|    Manhattan|       FL|          PTPT2002|54342-20
16-ENFO|   9|2016-07-04T00:00:00|   BROADWAY|   NY|       25000|                |             42|     null|            0|           |        |
   null|
|           1.0| 7001.0|NATIONAL FENCE SY...|NYC Admin Code \u...|   AVENEL|       0|Home Improvement ...|Staten Island|         |          05397536|32644-20
16-ENFO|    |2016-05-14T00:00:00|US HIGHWAY 1|   NJ|           1|         Pleaded|           1033|        0|            0|           |        |
      0|
|           1.0|33614.0|       PROVEST LLC|6 RCNY \u00A7 2-2...|    TAMPA|       0|Process Server (O...|  Outside NYC|         |          05401674|  400-20
16-LEGL|    |2016-08-31T00:00:00|SEEDLING CIR|   FL|           1|Settlement (Pre-H...|           4520|        0|            0|           |        |
      0|
|           1.0|33614.0|       PROVEST LLC|6 RCNY \u00A7 2-2...|    TAMPA|       0|Process Server (O...|  Outside NYC|         |          05401674|  400-20
16-LEGL|    |2016-08-31T00:00:00|SEEDLING CIR|   FL|           1|Settlement (Pre-H...|           4520|        0|            0|           |        |
      0|
|           1.0|17543.0|      MILPORT DAIRY|6 RCNY \u00A7 4-1...|   LITITZ|       0|Mobile Food Vendo...|     BROOKLYN|         |          05379442|34566-20
15-ENFO|    |2015-06-18T00:00:00|  BEILER LN|   PA|           1|         Pleaded|             30|        0|            0|           |        |
      0|
|           1.0|33614.0|       PROVEST LLC|6 RCNY \u00A7 2-2...|    TAMPA|       0|Process Server (O...|  Outside NYC|         |          05401674|  400-20
16-LEGL|    |2016-08-31T00:00:00|SEEDLING CIR|   FL|           1|Settlement (Pre-H...|           4520|        0|            0|           |        |
      0|
+--------------+-------+------------------+-----------------+---------+--------+-----------------+-----------+---------+------------------+--------
-------+----+------------------+-----------+-----+------------+----------------+---------------+---------+-------------+-----------+--------+----------
-------+
```

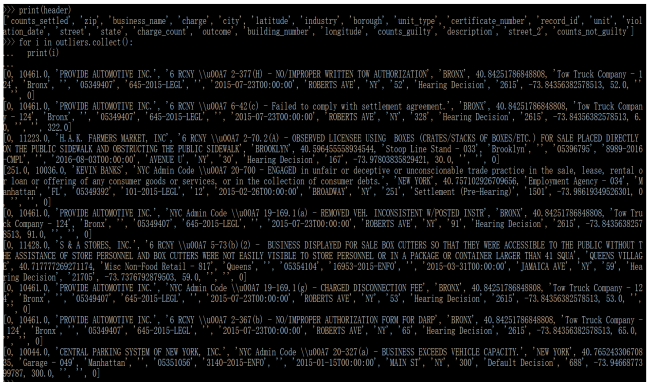
Figure 5: k-mean result

The AVF algorithm runs extremely fast on our dataset. However we can notice that it is difficult to decide which attributes contributes the most to our result according to the nature of the algorithm.

The K-means algorithm also worked decently on the dataset as the number of numerical data columns and the size of the dataset are at sweet spot. However, since the scale of zip code is much larger than the other columns, it played a much larger role in determining the outliers.

4.2 Improved Kmeans

To achieve better performance, we also implemented improved K means algorithm. We preprocessed numerical features to normalize to mean 1 to balance the impact due to scale difference of features and reduce the excessive dimensionality by evaluating correlation among features and keep only principal ones. We got result as following:



```
>>> print(header)
['counts_settled', 'zip', 'business_name', 'charge', 'city', 'latitude', 'industry', 'borough', 'unit_type', 'certificate_number', 'record_id', 'unit', 'viol
ation_date', 'street', 'state', 'charge_count', 'outcome', 'building_number', 'longitude', 'counts_guilty', 'description', 'street_2', 'counts_not_guilty']
>>> for i in outliers.collect():
...   print(i)
...
[0, 10461.0, 'PROVIDE AUTOMOTIVE INC.', '6 RCNY \\u00A7 2-377(H) - NO/IMPROPER WRITTEN TOW AUTHORIZATION', 'BRONX', 40.84251786848808, 'Tow Truck Company - 1
24', 'Bronx', '', '05349407', '645-2015-LEGL', '', '2015-07-23T00:00:00', 'ROBERTS AVE', 'NY', '52', 'Hearing Decision', '2615', -73.8435638257851З, 52.0, ''
, '', 0]
[0, 10461.0, 'PROVIDE AUTOMOTIVE INC.', '6 RCNY \\u00A7 6-42(c) - Failed to comply with settlement agreement.', 'BRONX', 40.84251786848808, 'Tow Truck Compan
y - 124', 'Bronx', '', '05349407', '645-2015-LEGL', '', '2015-07-23T00:00:00', 'ROBERTS AVE', 'NY', '328', 'Hearing Decision', '2615', -73.84356382578513, 6.
0, '', '', 322.0]
[0, 11223.0, 'H.A.K. FARMERS MARKET, INC', '6 RCNY \\u00A7 2-70.2(A) - OBSERVED LICENSEE USING  BOXES (CRATES/STACKS OF BOXES/ETC.) FOR SALE PLACED DIRECTLY
ON THE PUBLIC SIDEWALK AND OBSTRUCTING THE PUBLIC SIDEWALK', 'BROOKLYN', 40.596455558934544, 'Stoop Line Stand - 033', 'Brooklyn', '', '05396795', '8989-2016
-CMPL', '', '2016-08-03T00:00:00', 'AVENUE U', 'NY', '30', 'Hearing Decision', '167', -73.97803835829421, 30.0, '', '', 0]
[251.0, 10036.0, 'KEVIN BANKS', 'NYC Admin Code \\u00A7 20-700 - ENGAGED in unfair or deceptive or unconscionable trade practice in the sale, lease, rental o
r loan or offering of any consumer goods or services, or in the collection of consumer debts.', 'NEW YORK', 40.757102926709656, 'Employment Agency - 034', 'M
anhattan', 'FL', '05349392', '101-2015-LEGL', '12', '2015-02-26T00:00:00', 'BROADWAY', 'NY', '251', 'Settlement (Pre-Hearing)', '1501', -73.98619349526301, 0
, '', '', 0]
[0, 10461.0, 'PROVIDE AUTOMOTIVE INC.', 'NYC Admin Code \\u00A7 19-169.1(a) - REMOVED VEH. INCONSISTENT W/POSTED INSTR', 'BRONX', 40.84251786848808, 'Tow Tru
ck Company - 124', 'Bronx', '', '05349407', '645-2015-LEGL', '', '2015-07-23T00:00:00', 'ROBERTS AVE', 'NY', '91', 'Hearing Decision', '2615', -73.8435638257
8513, 91.0, '', '', 0]
[0, 11428.0, 'S & A STORES, INC.', '6 RCNY \\u00A7 5-73(b)(2) -  BUSINESS DISPLAYED FOR SALE BOX CUTTERS SO THAT THEY WERE ACCESSIBLE TO THE PUBLIC WITHOUT T
HE ASSISTANCE OF STORE PERSONNEL AND BOX CUTTERS WERE NOT EASILY VISIBLE TO STORE PERSONNEL OR IN A PACKAGE OR CONTAINER LARGER THAN 41 SQUA', 'QUEENS VILLAG
E', 40.717777269271174, 'Misc Non-Food Retail - 817', 'Queens', '', '05354104', '16953-2015-ENFO', '', '2015-03-31T00:00:00', 'JAMAICA AVE', 'NY', '59', 'Hea
ring Decision', '21705', -73.7376792879503, 59.0, '', '', 0]
[0, 10461.0, 'PROVIDE AUTOMOTIVE INC.', 'NYC Admin Code \\u00A7 19-169.1(g) - CHARGED DISCONNECTION FEE', 'BRONX', 40.84251786848808, 'Tow Truck Company - 12
4', 'Bronx', '', '05349407', '645-2015-LEGL', '', '2015-07-23T00:00:00', 'ROBERTS AVE', 'NY', '53', 'Hearing Decision', '2615', -73.84356382578513, 53.0, '',
'', 0]
[0, 10461.0, 'PROVIDE AUTOMOTIVE INC.', '6 RCNY \\u00A7 2-367(b) - NO/IMPROPER AUTHORIZATION FORM FOR DARP', 'BRONX', 40.84251786848808, 'Tow Truck Company -
 124', 'Bronx', '', '05349407', '645-2015-LEGL', '', '2015-07-23T00:00:00', 'ROBERTS AVE', 'NY', '65', 'Hearing Decision', '2615', -73.84356382578513, 65.0,
'', '', 0]
[0, 10044.0, 'CENTRAL PARKING SYSTEM OF NEW YORK, INC.', 'NYC Admin Code \\u00A7 20-327(a) - BUSINESS EXCEEDS VEHICLE CAPACITY.', 'NEW YORK', 40.765243306708
35, 'Garage - 049', 'Manhattan', '', '05351056', '3140-2015-ENFO', '', '2015-01-15T00:00:00', 'MAIN ST', 'NY', '300', 'Default Decision', '688', -73.94668773
99787, 300.0, '', '', 0]
```

Figure 6: improved k-mean result

Clearly, we can see that the improved algorithm find outliers that make more sense as the impact from zip code was reduced and other features contributed more to the selected data points.

5 TECHNICAL DEPTH AND INNOVATION

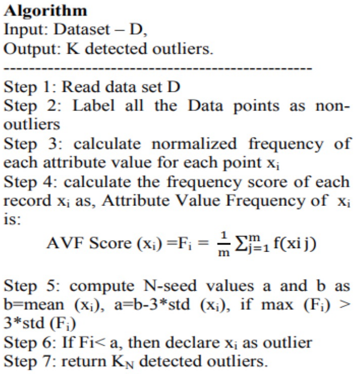
We implemented the most popular and quite powerful algorithms used in outlier detection for both numerical and categorical data. However, each of the algorithms has some drawbacks. From this point of view, our future work depends on how to improve these algorithms in terms of accuracy and time complexity.

5.1 Categorical data

Normally Distributed Attribute value Frequency(NAVF) method This model is defined as an optimal number of outliers in a single instance to get optimal precision in any classification model with good precision. This method calculates 'k'value itself based on the frequency instead of indicating 'k'manually. Taking the dataset 'D'with m attributes. k_N is the number of outliers which are normally distributed. If any object frequency is less than "mean-3 S.D" then this model treats those objects as outliers. This method uses AVF score formula to find AVF score but no k-value is required.

5.2 Numerical data

Since the k-means algorithm is extremely sensitive to outliers, there are some methods proposed to solve the problem by removing the outliers from the dataset before implementing k-means, or



Algorithm
Input: Dataset – D,
Output: K detected outliers.

Step 1: Read data set D
Step 2: Label all the Data points as non-outliers
Step 3: calculate normalized frequency of each attribute value for each point x_i
Step 4: calculate the frequency score of each record x_i as, Attribute Value Frequency of x_i is:

$$\text{AVF Score } (x_i) = F_i = \frac{1}{m} \sum_{j=1}^{m} f(xi\,j)$$

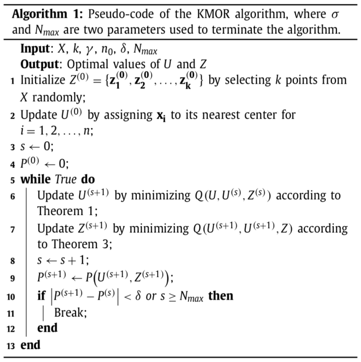
Step 5: compute N-seed values a and b as b=mean (x_i), a=b-3*std (x_i), if max (F_i) > 3*std (F_i)
Step 6: If Fi< a, then declare x_i as outlier
Step 7: return K_N detected outliers.

Figure 7: The NAVF Algorithm

by choosing initial cluster centers that are not outliers, or doing clustering and outlier detection simultaneously.

Improved K-means with Outlier Removal(KMOR) method This method is designed to improve the k-means algorithm performance through reducing unnecessary dimensionality and balance scale difference between features and outlier detection simultaneously.

In the KMOR algorithm, the numerical features are all normalized to mean 1 to balance the impact due to scale difference of features and reduce the excessive dimensionality by evaluating correlation among features and keep only principal ones. The given the clustering centers, a data point that is at least $\gamma \times d_{avg}$ away from all the cluster centers is considered as an outlier, where γ is a multiplier and d_{avg} is the average distance calculated dynamically during the clustering process. The KMOR algorithm has two main parameters: n_0 and γ. The parameter n0 specifies the maximum number of outliers. The parameter γ specifies the multiplier of the average squared distance for outlier detection.



Algorithm 1: Pseudo-code of the KMOR algorithm, where σ and N_{max} are two parameters used to terminate the algorithm.

```
Input: $X$, $k$, $\gamma$, $n_0$, $\delta$, $N_{max}$
Output: Optimal values of $U$ and $Z$
Initialize $Z^{(0)} = \{\mathbf{z}_1^{(0)}, \mathbf{z}_2^{(0)}, \ldots, \mathbf{z}_k^{(0)}\}$ by selecting $k$ points from $X$ randomly;
Update $U^{(0)}$ by assigning $\mathbf{x}_i$ to its nearest center for $i = 1, 2, \ldots, n$;
$s \leftarrow 0$;
$P^{(0)} \leftarrow 0$;
while True do
    Update $U^{(s+1)}$ by minimizing $Q(U, U^{(s)}, Z^{(s)})$ according to Theorem 1;
    Update $Z^{(s+1)}$ by minimizing $Q(U^{(s+1)}, U^{(s+1)}, Z)$ according to Theorem 3;
    $s \leftarrow s + 1$;
    $P^{(s+1)} \leftarrow P\big(U^{(s+1)}, Z^{(s+1)}\big)$;
    if $\left|P^{(s+1)} - P^{(s)}\right| < \delta$ or $s \geq N_{max}$ then
        Break;
    end
end
```

Figure 8: The KMOR Algorithm

6 CODE REPOSITORY

Repository URL: https://github.com/KLiunyu/DS1004_Project It includes our detection code, and also a README.txt file for instruction on running the outliers detection code.

REFERENCES

[1] Charu C Aggarwal and Chandan K Reddy. 2013. *Data clustering: algorithms and applications.* CRC press.

[2] Vic Barnett and Toby Lewis. 1994. *Outliers in Statistical Data* (3rd ed.). John Wiley and Sons.

[3] Guojun Gan and Michael Kwok-Po Ng. 2017. k-means clustering with outlier removal. *Pattern Recognition Letters* 90 (2017), 8–14.

[4] Frank E Grubbs. 1969. Procedures for detecting outlying observations in samples. *Technometrics* 11, 1 (February 1969), 1–21.

[5] Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22, 2 (October 2004), 85–126.