2-2021

# Forensic analytics using cluster analysis: Detecting anomalies in data

Clarence GOH
*Singapore Management University*, clarencegoh@smu.edu.sg

Benjamin Huan Zhou LEE
*Singapore Management University*, benjaminlee@smu.edu.sg

Gary PAN
*Singapore Management University*, garypan@smu.edu.sg

Poh-Sun Seow
*Singapore Management University*, psseow@smu.edu.sg

# Forensic Analytics Using Cluster Analysis: Detecting Anomalies in Data

Clarence Goh
clarencegoh@smu.edu.sg

Benjamin Lee
benjaminlee@smu.edu.sg

Gary Pan
garypan@smu.edu.sg

Seow Poh Sun
psseow@smu.edu.sg

Singapore Management University
60 Stamford Road
Singapore 178900

October 2020

# Forensic Analytics Using Cluster Analysis: Detecting Anomalies in Data

**ABSTRACT**:

Cluster analysis is a data analytics technique that can help forensic accountants effectively detect anomalies in complex financial datasets. This article provides a description of clustering analysis, discusses how it can be implemented to detect anomalies in data, and illustrates its use through a worked example using the Tableau software.

Keywords: Forensic accounting, data analytics, clustering, Tableau

## Forensic Analytics Using Cluster Analysis: Detecting Anomalies in Data

### Introduction

Forensic accounting refers to the branch of accounting that deals with the application of accounting facts gathered through auditing methods and procedures to resolve legal problems (Bhasin 2007). It involves the integration of investigative, accounting, and auditing skills. In performing forensic investigations, forensic accountants need to calculate values, identify irregular patterns or suspicious transactions, and draw conclusions by critically analyzing financial data. Forensic accounting often also provides an accounting analysis and explanation for frauds that have been committed (Koh et al. 2009). Given the growth in financial crime rates, there has been an increasing focus on the role and skillsets that forensic accountants should possess (Owojori and Asaolu 2009). Consistent with this, a recent survey conducted by Rezaee et al. (2004) found that 93.3% of academics and 88.2% of practitioners expect the demand and interest for services in fraud examination to increase in the near future.

Given the rapid advancement in technology and the growth in both the quantity and variety of data available in the corporate setting, the use of data analytics techniques have become widespread in accounting, including in the area of forensic accounting (Pan et al. 2017; Nigrini 2020). One important data analytics technique in forensic accounting is the use of cluster analysis to detect anomalies in financial data (Chandola et al. 2009). In this article, we introduce the cluster analysis technique and highlight how it can be applied in anomaly detection in forensic accounting. Using a worked example, we also demonstrate how cluster analysis can be implemented using the Tableau software.

Tableau is a software application that queries data and generates graph-like visualizations. It is capable of handling large datasets and can also perform various data

analytics procedures, including clustering analysis. Although Tableau is a useful tool that is effective in performing forensic analytics tasks (Nigrini 2020), a recent survey conducted by Ernst & Young (2014) highlights that only 12% of respondents regularly use it in forensic data analytics, with the majority of respondents (65%) preferring to rely on less sophisticated tools such as Microsoft Excel. Our focus on Tableau in this article contributes to the practice of forensic accounting by highlighting how an effective but less commonly used tool can be applied in forensic analytics.

**Basics of Cluster Analysis**

Anomalies in data refer to outlier observations that deviate so much from other observations that they may have been generated by a different mechanism (Hawkins 1980). Anomaly detection techniques are effective in forensic accounting because they can effectively identify unusual corporate financial behavior (Sharma and Panigrahi 2012). Cluster analysis can be applied to detect anomalies in complex datasets. Cluster analysis involves classifying data in such a way that data assigned to the same cluster are more similar to one another (along dimensions being examined) than to data assigned to another cluster (Everitt et al. 2011). Prior studies suggest that that in performing cluster analysis, normal data is assigned to clusters which are large and dense while anomalies are assigned to clusters which are small and/or sparse (e.g. Chandola et al. 2009; Sun et al. 2004).

Tableau performs cluster analysis using the K-means clustering algorithm (Everitt et. al. 2011). K-means clustering has been examined in the accounting literature as a technique in forensic accounting (Amani and Fadlalla 2017). Tang et al. (2006) suggest that using the K-means algorithm is suitable when the dataset under analysis is large, as often is the case with financial data, because it requires less computing power than other methods. Several steps are involved in the K-means clustering algorithm. In step one, a value for K, which

represents the number of clusters to be used in the analysis, is assigned (by the forensic accountant). The value assigned to K should be selected such that adding an additional cluster (i.e. K+1) does not significantly improve the cluster model. In step two, the algorithm randomly selects K data points. These data points represent initial cluster centers, which serve as a prototype of cluster members. In step three, the algorithm computes Euclidean distances and uses it to assign the remaining data to their closest cluster centers. Euclidean distance is a measurement of the distance between a pair of data points, and is computed by taking the square root of squared difference between the coordinates of a pair of data points (Singh et al. 2013). In step four, using the data assigned to each cluster, a new cluster center is determined. In step five, if the new cluster center identified in step four is identical to the original cluster center, the process terminates. If not, steps three to five are repeated.

## Conducting Cluster Analysis on Financial Data

In this section, we illustrate how cluster analysis can be performed using Tableau to detect anomalies in financial data using a worked example.

### *Data*

We perform our analysis on a fictitious dataset containing claims data from a flexible spending account (FSA) of LuxeHoreca SG, a company headquartered in Singapore. Claims are a common area where fraud is perpetuated, and has been examined in the accounting literature (e.g. Morris 2009; Pflaum and Rivers 1991). Our dataset contains 36,772 FSA claims made from 2016 to 2018. FSA claims fall into three main categories: flexible benefits (*FLXI*), birthday (*BIRT*), medical insurance for individuals' scheme (*MIIS*).[1] Table 1 summarizes and provides a brief description of the data fields contained in the dataset.

[1] There are a further eight categories under FLXI: (1) IT equipment and accessories, (2) Childcare Expenses, (3)Personal Wellness (Vision Care/Fitness/Exercise), (4) Vacation Expenses, (5) Family Insurance Expenses (Spouse, Children, Parents), (6) Personal and Work-life Enrichment, (7) Telephone Subscriptions, (8) Professional Membership Fees. The dataset is available upon request from the authors.

*(Insert Table 1 here)*

The claims were made by 7,695 unique employees. Overall, 8,219 claims relate to *BIRT*, 20,013 claims relate to *FLXI*, and 8,540 claims relate to *MIIS*. In addition, 2,123 claims were made in 2016, 18,426 claims were made in 2017, and 16,223 claims were made in 2018. The mean (standard deviation) claim amount (*CLM_AMT*) is $187.42 (183.87).

**Analysis**

Two pertinent characteristics of the claims in our data relate to (i) the number of days between the date indicated on a receipt submitted with a claim and date that the claim was submitted and (ii) the number of days between the date that a claim was submitted and the date on which a reimbursement was made. In particular, the time period between the date indicated on a receipt and the date in which the corresponding claim was submitted should not be overly long, consistent with claims being submitted promptly after the expenses have been incurred. In addition, the time period between the date that a claim was submitted and the date on which a reimbursement was made should not be overly short or long, and should instead be consistent with a reasonable processing period. Accordingly, we began our analysis by creating calculated fields in Tableau to examine these two characteristics. Tableau allows the use of calculations to create such new fields, which are then saved as part of the data source. Calculated fields can be created in Tableau by going to Analysis and selecting Create Calculated Field. We created the *DAYSTO_CLAIM* calculated field by using the formula *[CLM_Dt]-[RCPT_DT]* to calculate, for each claim, the number of days between the date indicated on a receipt submitted with a claim and date that the claim was submitted. We also created the *DAYSTO_REIMBURSE* calculated field by using the formula *[REIMB_DT]-[CLM_DT]* to calculate, for each claim, the number of days between the date that a claim was submitted and the date on which a reimbursement was made. The mean

(standard deviation) for *DAYSTO_CLAIM* is 18.34 (10.27) days while the mean (standard deviation) for *DAYSTO_REIMBURSE* is 42.48 (10.68) days.

To further examine *DAYSTO_CLAIM* and *DAYSTO_REIMBURSE*, we created a scatterplot to show the spread of our data along these two characteristics. Figure 1 presents a screenshot of the scatterplot created in Tableau. It is created in a Tableau worksheet by (i) dragging and dropping *DAYTO_REIMBURSE* from the data tab to the Columns shelf and *DAYSTO_CLAIM* from the Data tab into the Rows shelf and (ii) clicking on "Analysis" and unselecting "Aggregate Measures."

*(Insert Figure 1 here)*

Next, we performed cluster analysis on the data along the following characteristics of claims: *DAYSTO_CLAIM*, *DAYSTO_REIMBURSE*, and *CLM_AMT*. We first created a visual on a Tableau worksheet by: (i) dragging and dropping *CLM_AMT* to the Columns shelf, and *DAYSTO_CLAIM* and *DAYSTO_REIMBURSE* to the Rows shelf, (ii) clicking on the *SUM[DAYSTO_REIMBURSE]* pill in the Rows shelf and selecting Dual Axis, and (iii) clicking on Analysis and unselecting Aggregate Measures. Figure 2 presents the visual that is created in Tableau.

*(Insert Figure 2 here)*

In Tableau, cluster analysis is performed by (i) dragging and dropping the Cluster field in the Analytics tab into the visualization canvas of a Tableau worksheet, (ii) dragging and dropping clustering variables into the Cluster menu area, and (iii) indicating the number of clusters to create (this is the 'K' in K-means). Figure 3 presents the cluster menu populated to perform cluster analysis by creating 8 clusters using *CLM_AMT*, *DAYTO_REIMBURSE*, and *DAYSTO_CLAIM*.

*(Insert Figure 3 here)*

Tableau further provides statistics associated with cluster analyses performed. These statistics can be viewed by clicking on the Cluster pill in the Marks area and selecting Describe Clusters and viewing the Summary tab. Table 2 presents the Summary Diagnostic statistics provided by Tableau for cluster analyses performed on 8, 9, 10, and 11 clusters. Number of clusters refers to the number of individual clusters used in the clustering analysis and number of points refers to the number of data points used in the analysis (Tableau 2020). In addition, between-group sum of squares is a metric quantifying the separation between clusters. It is computed by taking the sum of squared distances between each cluster's center, and weighing it by the number of data points assigned to the cluster and the center of the data set. Larger between-group sum of squares values indicate better separation between clusters. Within-group sum of squares is a metric quantifying the cohesion of clusters. It is computed by taking the sum of squared distances between the center of each cluster and the individual marks in the cluster. Smaller within-group sum of square values indicate that clusters are more cohesive. Total sum of squares is computed by summing the between-group sum of squares and the within-group sum of squares.

*(Insert Table 2 here)*

In examining these summary diagnostic statistics, the value of the between-group sum of squares as a proportion of the total sum of squares provides an indication of the proportion of variance in the data explained by the clustering model. Thiprungsri and Vasarhelyi (2011) suggest that the ideal number of clusters chosen should be at the point where the marginal gain in the proportion of variance in the data explained by a clustering model begins to fall as additional clusters are added. With reference to Table 2, we determine that this point occurs when 10 clusters are chosen. In particular, when 10 clusters are chosen, the percentage increase in the proportion of variance explained is 1.05% (from when 9 clusters are chosen); when 11 clusters are chosen, the percentage increase in the proportion of variance explained

(from when 10 clusters are chosen) falls to 0.77%. Accordingly, we focus our subsequent analysis on a 10-cluster model.

We first examined the analysis of variance statistics for the cluster analysis. These statistics can be viewed by clicking on the Cluster pill in the Marks area and selecting Describe Clusters and viewing the Models tab. Figure 4 presents these statistics, as summarized in Tableau. The F-statistic provides an indication of how well each variable distinguishes between clusters, with larger values suggesting that a variable is better at distinguishing between clusters (Tableau 2020; Smoak 2018). The corresponding p-value represents the probability that the F-distribution of all possible values of the F-statistic takes on a value greater than the actual F-statistic for a variable. The lower the p-value, the more the expected values of the elements of the corresponding variable differ among clusters. Overall, the F-statistics are large ($F>3080$) and p-values small ($p<0.01$) for each the three variables included in our cluster analysis, suggesting that they are each effective in distinguishing between clusters. The model sum of squares is the ratio of the between-group sum of squares to the model degrees of freedom. The model sum of squares for each of the three variables included in our cluster analysis are relatively large (model sum of squares>2000), suggesting that their cluster means are well spread out from each other.

*(Insert Figure 4 here)*

Next, we examine the clusters created in the cluster analysis. Figure 5 presents details relating to each of the clusters created, as summarized in Tableau. The number of data items in each cluster range from 1559 (cluster 10) to 6697 (cluster 2). The centroid centers for CLM_AMT range from $68.40 (cluster 9) to $481.96 (cluster 5). Further, the centroid centers for *DAYSTO_REIMBURSE* range from 27.70 days (cluster 4) to 53.29 days (cluster 6) while

the centroid clusters for *DAYSTO_CLAIM* range from 7.62 days (cluster 10) to 32.52 days (cluster 1).

*(Insert Figure 5 here)*

Prior studies suggest that, anomalies are assigned to clusters which are small and/or sparse while normal data is assigned to clusters which are large and dense while a (e.g. Chandola et al. 2009; Sun et al. 2004). Accordingly, we focus our attention on identifying clusters which are small and/or sparse. In particular, we identify cluster 10 as potentially containing anomalous data items. Cluster 10 is the smallest cluster formed by the cluster analysis, with only 1559 data items assigned to the cluster. These 1599 data items account for only 4.23% of data items in our dataset. In comparison, the 6697 data items in the largest cluster (cluster 2) account for 18.22% of data items in our dataset. Figure 6 presents a visualization of our cluster analysis, with cluster 10 highlighted in panel A and cluster 2 (our largest cluster) highlighted in panel B. We note that the data items assigned to cluster 10 are more highly spread out than the data items assigned to cluster 2, highlighting the relative sparseness of cluster 10.

*(Insert Figure 6 here)*

In examining cluster 10, we also find that the cluster centers indicate that it has a relatively low *DAYSTO_REIMBURSE* of 37.01 days (this is the fourth lowest among the 10 clusters), a relatively low *DAYSTO_CLAIM* of 7.62 days (this is the lowest among the 10 clusters), and a relatively high *CLM_AMT* of $467.64 (this is the fourth highest among the 10 clusters). However, even though we identify data items in cluster 10 as displaying anomalous characteristics, we note that this does not necessarily signify that they relate to fraudulent claims (Thiprungsri and Vasarhelyi 2011). Given that there may be legitimate reasons that account for these data items' characteristics, forensic accountants should further investigate

these data items using complementary forensic analysis techniques to ascertain the nature of the underlying claims. Further, where the number of anomalous claims are large, it may be impossible for a forensic investigator to investigate all claims. Instead, decisions on materiality should be made to determine which claims within a cluster should be pursued and which ones should be left aside (Cleary et al. 2005).

## Conclusion

Given improvements in technology and the growing prevalence of data availability, forensic accountants can effectively apply data analytics techniques to detect anomalies in complex datasets. One important data analytics technique in forensic accounting is the use of cluster analysis to detect anomalies in financial data (Chandola et al. 2009). In this article, we introduce the cluster analysis technique and discuss how it can be applied to detect anomalies in forensic accounting. Using a worked example, we further demonstrate how cluster analysis can be implemented to detect anomalies in claims data using the Tableau software. Our study contributes to the practice of forensic accounting by highlighting how a popular data analytics tool such as Tableau can be used to conduct cluster analysis and to detect anomalies in data.

## References


Amani, F. A. & Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, 24, 32-58.

Bhasin, M. L. (2007). Forensic accounting: A new paradigm for niche consulting. *The Chartered Accountant Journal*, 1000-1010.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 40 (3): 1-58.

Clearly, B., & Thibodeau, J. C. (2005). Applying digital analysis using Benford's law to detect fraud: The dangers of type I errors. *Auditing: A Journal of Practice and Theory*, 24 91): 77-81.

Ernst & Young. (2014). Big risks require big data thinking: Global forensic data analytics survey 2014. Available here: https://www.ey.com/Publication/vwLUAssets/EY-Global-Forensic-Data-Analytics-Survey-2014/$FILE/EY-Global-Forensic-Data-Analytics-Survey-2014.pdf (Accessed: September 25 2020).

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. 2011. Cluster analysis. John Wiley & Sons, West Sussex, U. K.

Hawkins, D. (1980). Identification of outliers. Chapman and Hall, London.

Koh, A. N., Arokiasamy, L., & Lee, C. (2009). Forensic accounting: Public acceptance towards occurrence of fraud detection. *International Journal of Business and Management*, 4 (11), 145-149.

Morris, L. (2009). Combating fraud in health care: an essential component of any cost containment strategy. *Health Affairs*, 28 (5), 1351-1356.

Nigrini, M. (2020). Forensic analytics: Methods and techniques for forensic accounting investigations. John Wiley & Sons, Hoboken: New Jersey.

Owojori, A. A., & Asaolu, T. O. (2009). The role of forensic accounting in solving the vexed problem of corporate world. *European journal of scientific research*, 29(2), 183-187.

Pan, G., Seow, P. S., Goh, C., & Yong, M. (2017). Riding the waves of disruption. CPA Australia, Singapore.

Pflaum, B. B., & Rivers, J. S. (1991). Employer strategies to combat health care plan fraud. Benefits quarterly 7 (1), 6-14.

Rezaee, Z., Crumbley, D. L., & Elmore, R. C. (2004). Forensic accounting education: A survey of academicians and practitioners. *Advances in Accounting Education*, 6, 193-231.

Sharma, A., & Panigrahi, P. K. (2012). A review of financial accounting fraud detection based on data mining techniques. International Journal of Computer Applications, 39 (1), 37-47.

Smoak, A. (2018). Using clustering analysis in Tableau to uncover the inherent patterns in your data. Available here: https://anthonysmoak.com/2018/09/30/use-clustering-analysis-in-tableau-to-uncover-the-inherent-patterns-in-your-data/ (Accessed: September 25 2020).

Sun, H., Bao, Y., Zhao, F., Yu, G., & Wang, D. (2004). "Cd-trees: An efficient index structure for outlier detection. *Lecture Notes in Computer Science*, 3129, 600-609.

Tableau. (2020). Find clusters in data. Available here: https://help.tableau.com/current/pro/desktop/en-us/clustering.htm#information-on-statistics-models-used-for-clusters (Accessed: September 25 2020).

Tang, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. Pearson Education.

Thiprungsri, S. & Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. *The International Journal of Digital Accounting Research*, 11, 69-84.

**Figure 1: Scatterplot of Data in Tableau**

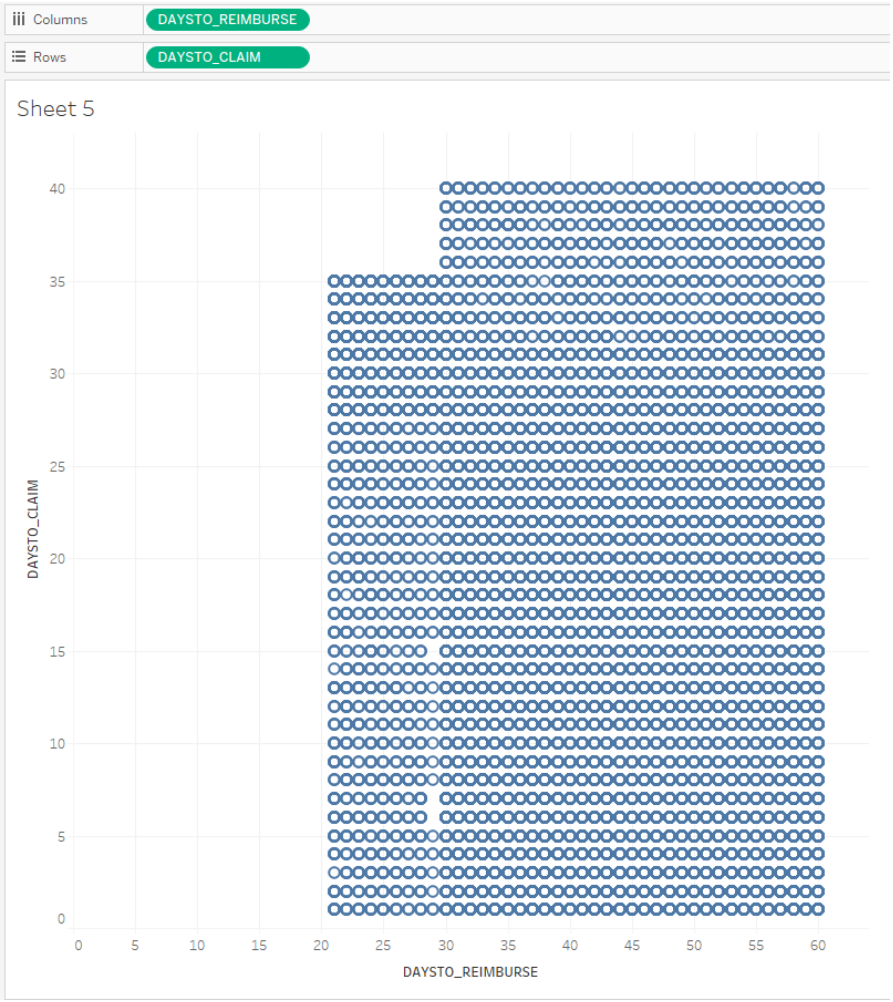


**Figure 2: Visualization of *DAYSTO_CLAIM*, *DAYSTOREIMBURSE* and *CLM_AMT* in Tableau**

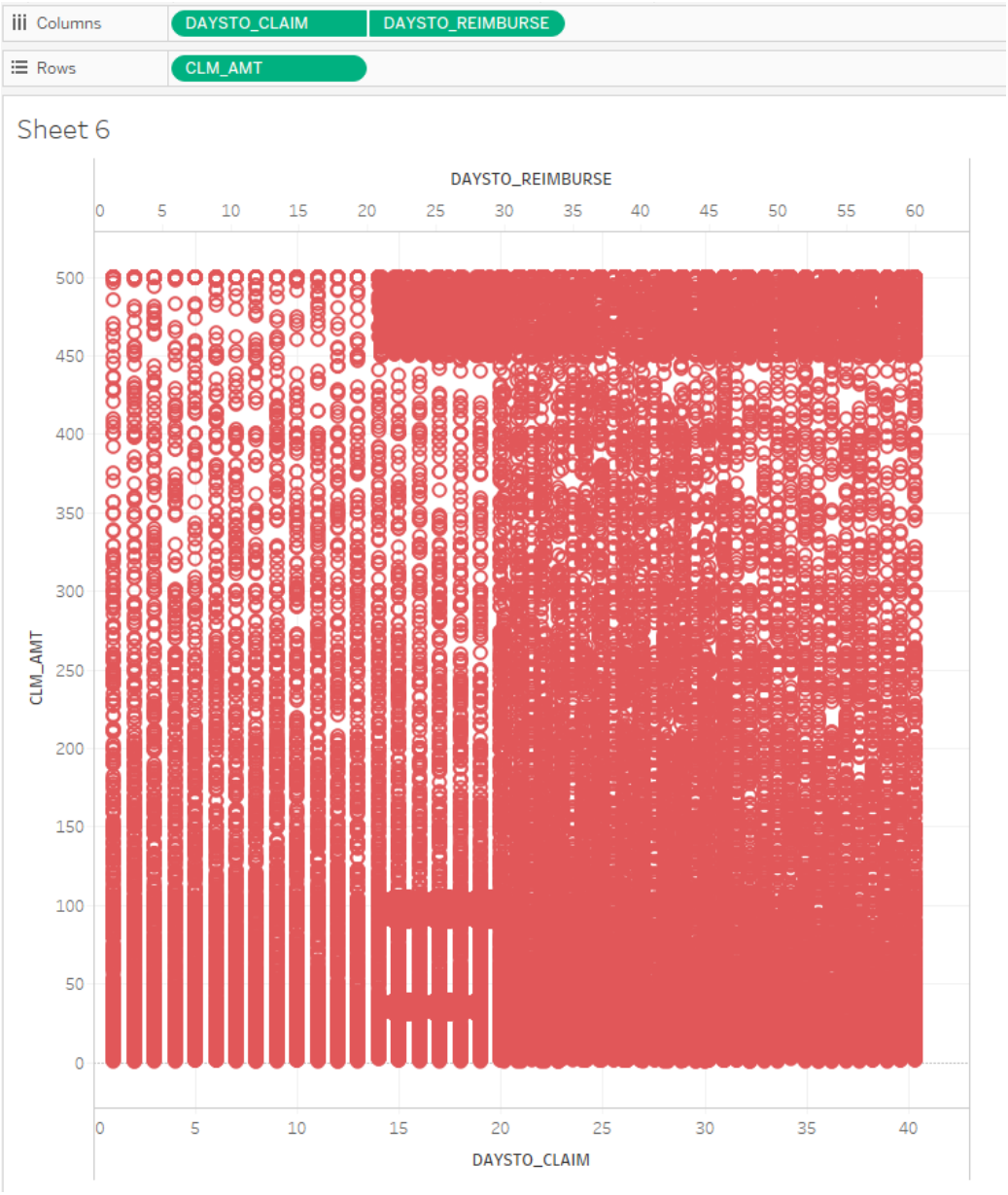


**Figure 3: Clustering Menu in Tableau**

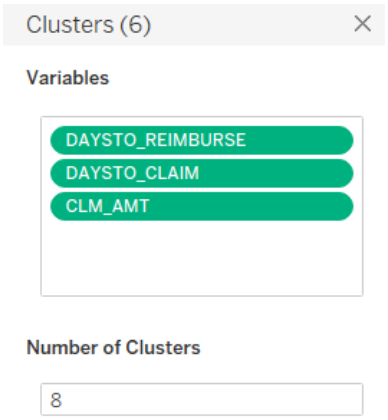
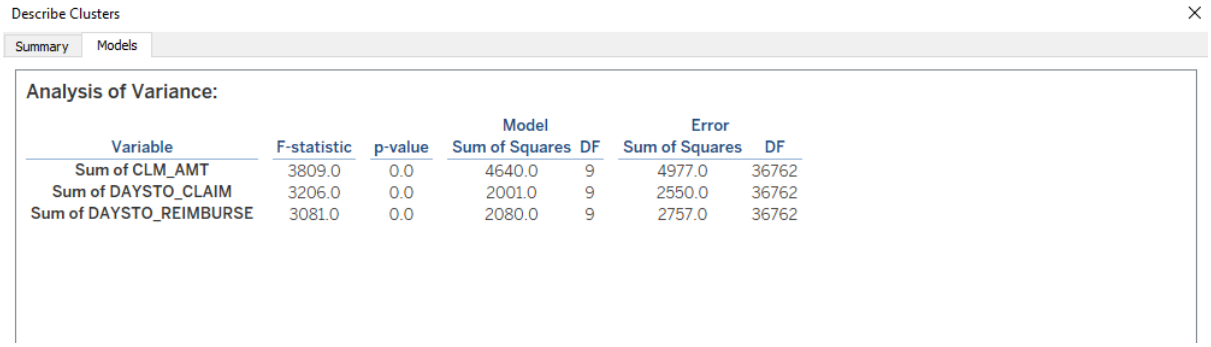



**Figure 4: Analysis of Variance in Tableau**




| | | | Model | | Error | |
|---|---|---|---|---|---|---|
| Variable | F-statistic | p-value | Sum of Squares | DF | Sum of Squares | DF |
| Sum of CLM_AMT | 3809.0 | 0.0 | 4640.0 | 9 | 4977.0 | 36762 |
| Sum of DAYSTO_CLAIM | 3206.0 | 0.0 | 2001.0 | 9 | 2550.0 | 36762 |
| Sum of DAYSTO_REIMBURSE | 3081.0 | 0.0 | 2080.0 | 9 | 2757.0 | 36762 |

**Figure 5: Cluster Details in Tableau**

| | | Centers | | |
|---|---|---|---|---|
| Clusters | Number of Items | Sum of DAYSTO_REIMBURSE | Sum of DAYSTO_CLAIM | Sum of CLM_AMT |
| Cluster 1 | 2067 | 51.673 | 32.515 | 473.84 |
| Cluster 2 | 6697 | 51.275 | 23.739 | 67.364 |
| Cluster 3 | 1861 | 44.667 | 17.586 | 249.95 |
| Cluster 4 | 2677 | 27.695 | 32.252 | 473.61 |
| Cluster 5 | 1590 | 39.138 | 23.77 | 481.96 |
| Cluster 6 | 1915 | 53.286 | 10.881 | 476.28 |
| Cluster 7 | 6108 | 52.3 | 8.0219 | 72.586 |
| Cluster 8 | 5839 | 31.918 | 26.111 | 74.584 |
| Cluster 9 | 6459 | 35.131 | 8.8706 | 68.401 |
| Cluster 10 | 1559 | 37.012 | 7.6241 | 467.64 |
| Not Clustered | 0 | | | |

**Figure 6: Cluster Analysis in Tableau**

**Panel A: Cluster 10 Highlighted**

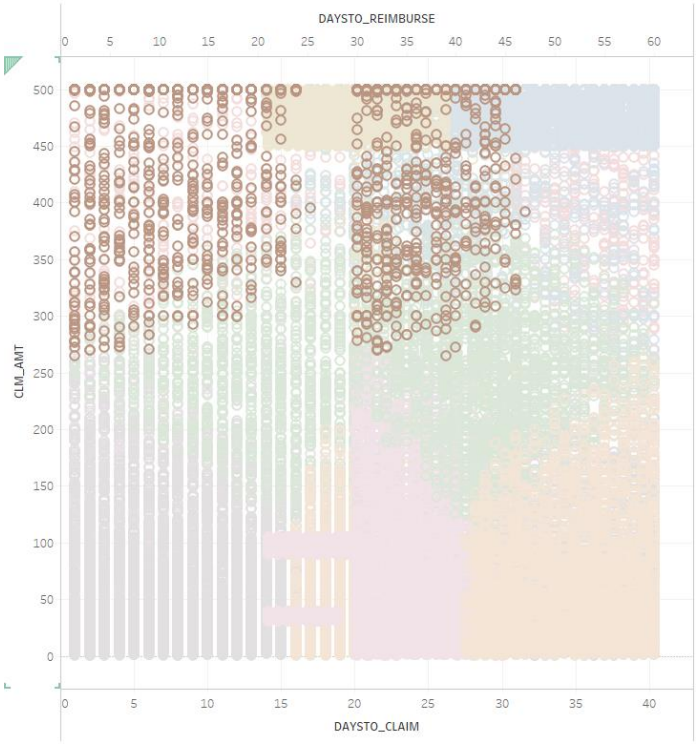


**Figure 6: Cluster Analysis in Tableau (continued)**

## Panel A: Cluster 2 Highlighted

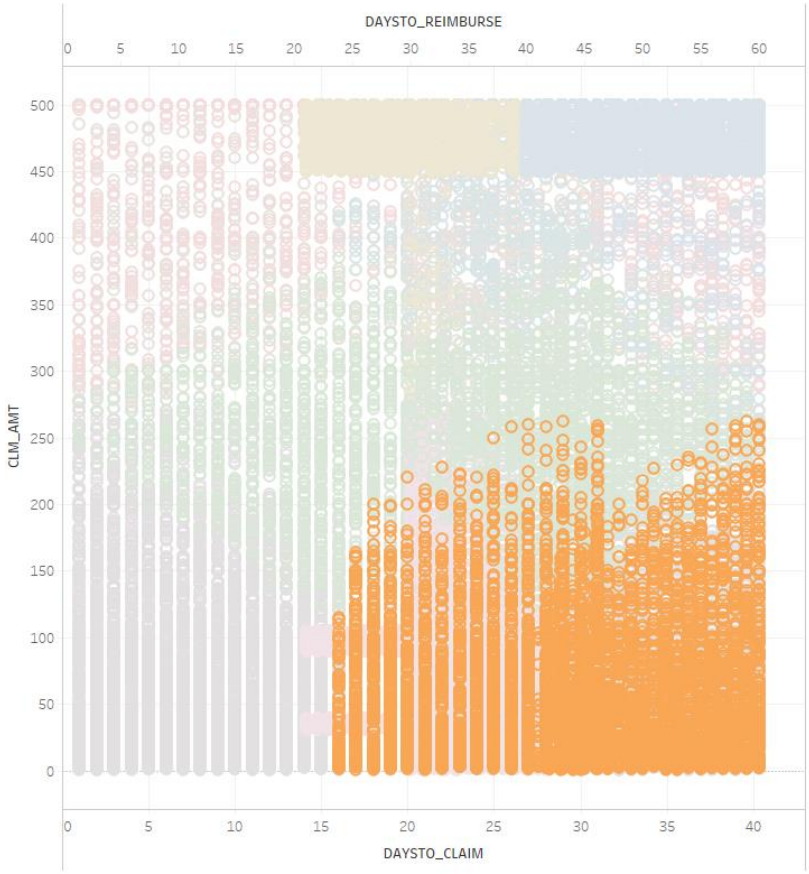


**Table 1: Summary and Description of Data Fields**

| Data Field Name | Description |
|---|---|
| EMP_ID | Unique identifier of each employee |
| CLM_SYS | Unique identifier of claim system |
| CLM_REF | Unique identifier of each claim submitted |
| FSA_TYP | Unique identifier of FSA claim category |
| FLEXBEN_TYPE | Unique identifier of FLXI category (refer to footnote 1 for detailed category list) |
| CLM_DT | Date on which claim was submitted |
| CLM_YR | Year in which claim was submitted |
| CLM_AMT | Claim amount |
| RCPT_DT | Date indicated on receipt submitted with claim |
| RCPT_DAY | Day indicated on receipt submitted with claim |
| DAY_TAG | Tag indicating whether the day indicated on receipt submitted with claim is a weekday or a weekend |
| REIMB_DT | Date on which reimbursement was made |
| REIMB_YR | Year in which reimbursement was made |

**Table 2: Summary Diagnostic Statistics for Clustering Analysis**

| | Number of Clusters | Number of Points | Between-Group Sum of Squares | Within-Group Sum of Squares | Total Sum of Squares | Percentage of Variance Explained | Percentage Gain |
|---|---|---|---|---|---|---|---|
| 1 | 8 | 36,772 | 8,542.90 | 1,741.20 | 10,284.00 | 83.07% | |
| 2 | 9 | 36,772 | 8,613.80 | 1,670.30 | 10,284.00 | 83.76% | 0.69% |
| 3 | 10 | 36,772 | 8,722.10 | 1,562.00 | 10,284.00 | 84.81% | 1.05% |
| 4 | 11 | 36,772 | 8,800.90 | 1,483.10 | 10,284.00 | 85.58% | 0.77% |