Clustering of Exceptions as an Outlier Detection Technique

Kristyn Calabrese, CPA/PhD Candidate

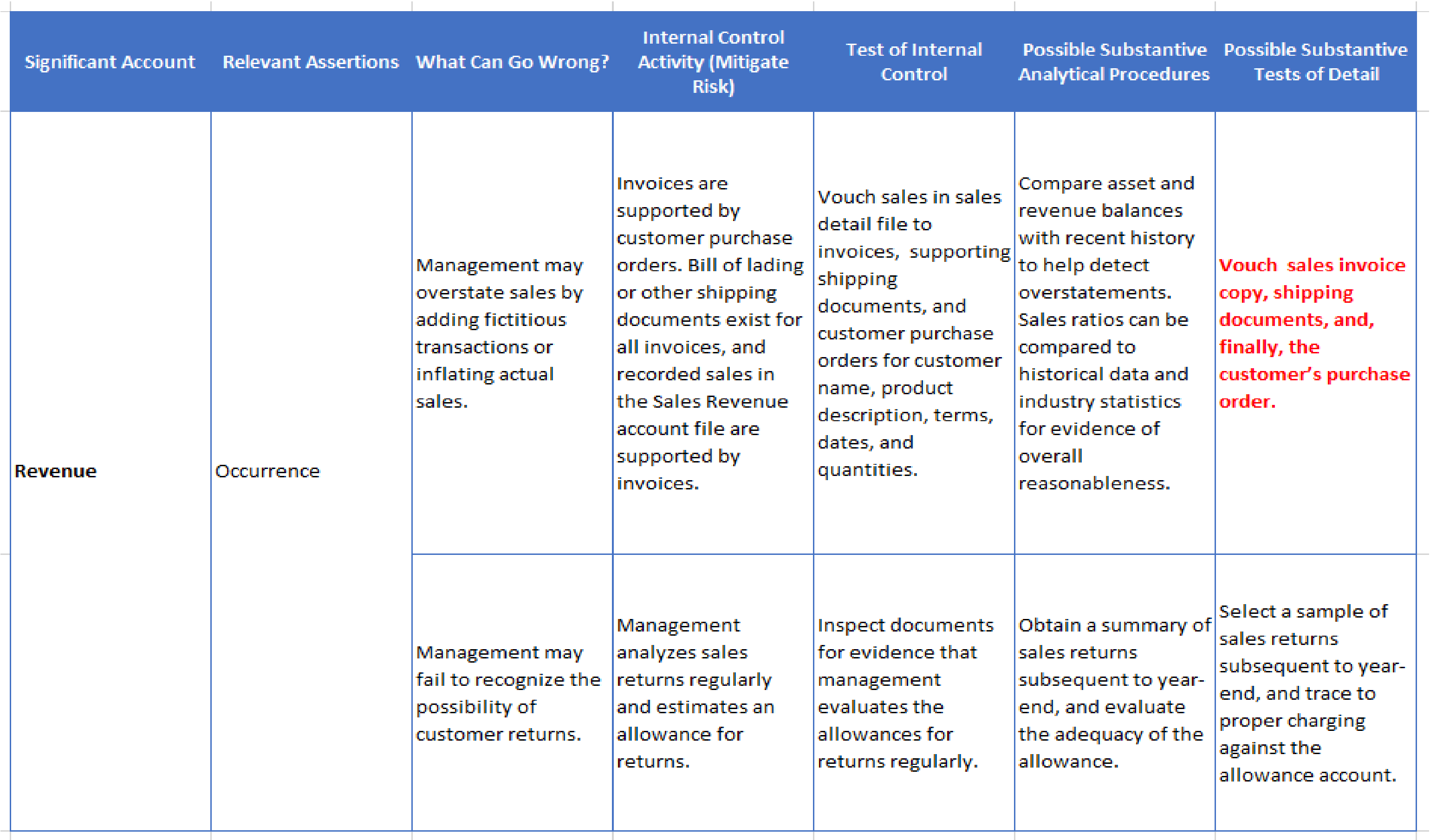
Rutgers Business School

Walmart Case Study Continued

What Can Go Wrong in the Revenue and Collection Cycle?

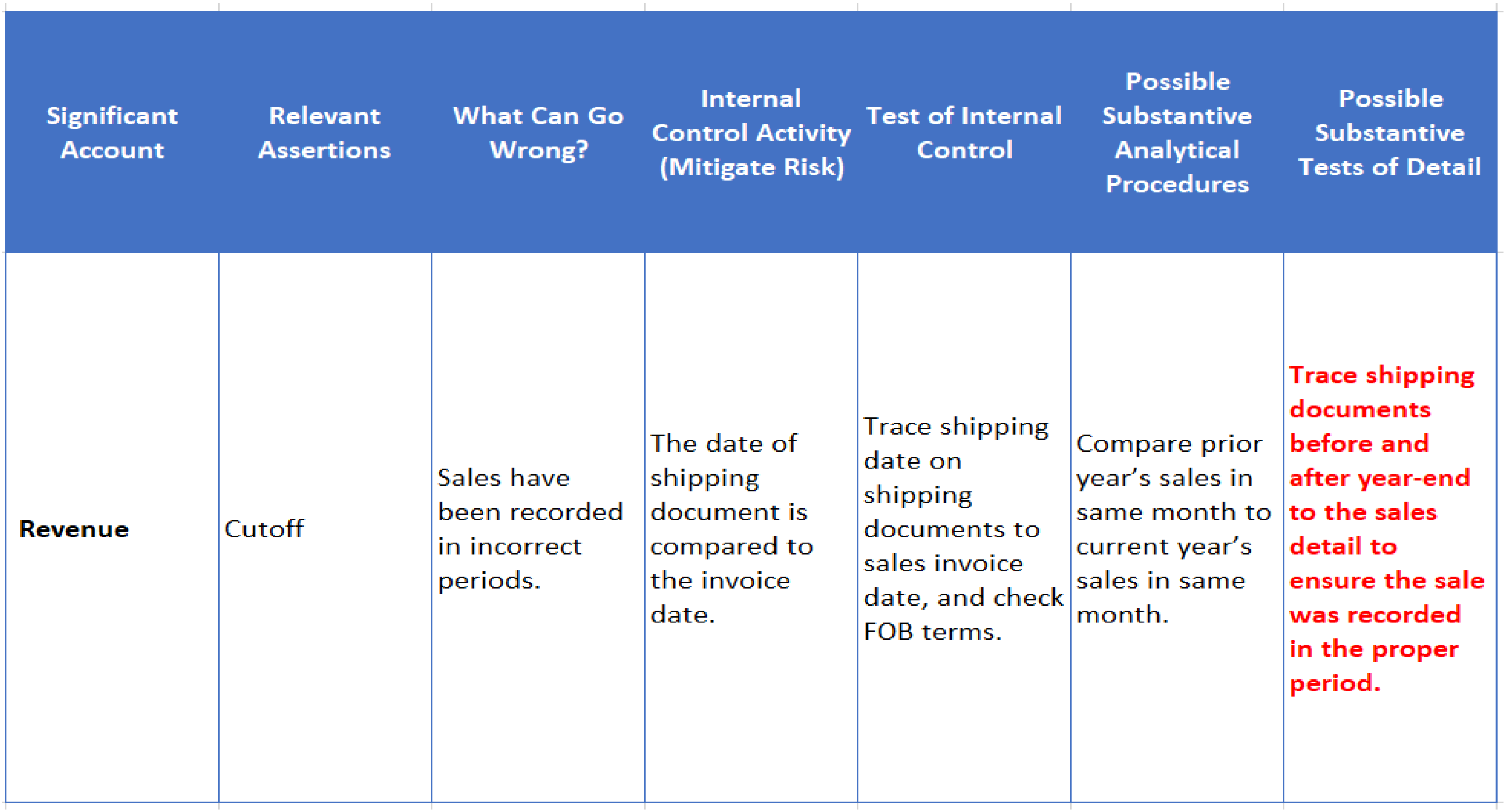
Significa-nt Account	Relevant Assertions	What Can Go Wrong?
Revenue	Occurrence	Management may overstate sales by adding fictitious transactions or inflating actual sales. Management may fail to recognize the possibility of customer returns.
Revenue	Completeness	Not all sales are recorded.
Revenue	Cutoff	Sales have been recorded in incorrect periods.

Selective Substantive Audit Procedures



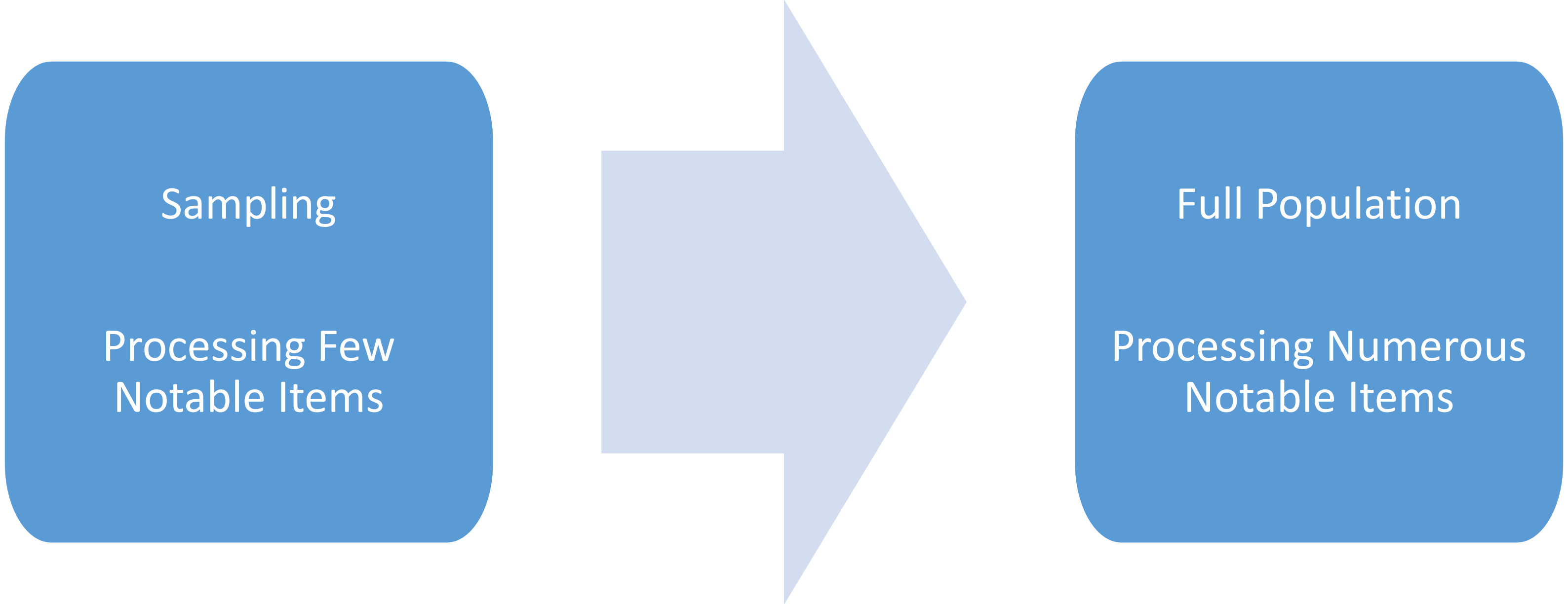
Significant Account	Relevant Assertions	What Can Go Wrong?	Internal Control Activity (Mitigate Risk)	Test of Internal Control	Possible Substantive Analytical Procedures	Possible Substantive Tests of Detail
Revenue	Occurrence	Management may overstate sales by adding fictitious transactions or inflating actual sales.	Invoices are supported by customer purchase orders. Bill of lading or other shipping documents exist for all invoices, and recorded sales in the Sales Revenue account file are supported by invoices.	Vouch sales in sales detail file to invoices, supporting shipping documents, and customer purchase orders for customer name, product description, terms, dates, and quantities.	Compare asset and revenue balances with recent history to help detect overstatements. Sales ratios can be compared to historical data and industry statistics for evidence of overall reasonableness.	**Vouch sales invoice copy, shipping documents, and, finally, the customer's purchase order.**
		Management may fail to recognize the possibility of customer returns.	Management analyzes sales returns regularly and estimates an allowance for returns.	Inspect documents for evidence that management evaluates the allowances for returns regularly.	Obtain a summary of sales returns subsequent to year-end, and evaluate the adequacy of the allowance.	Select a sample of sales returns subsequent to year-end, and trace to proper charging against the allowance account.

Selective Substantive Audit Procedures



Significant Account	Relevant Assertions	What Can Go Wrong?	Internal Control Activity (Mitigate Risk)	Test of Internal Control	Possible Substantive Analytical Procedures	Possible Substantive Tests of Detail
Revenue	Cutoff	Sales have been recorded in incorrect periods.	The date of shipping document is compared to the invoice date.	Trace shipping date on shipping documents to sales invoice date, and check FOB terms.	Compare prior year's sales in same month to current year's sales in same month.	**Trace shipping documents before and after year-end to the sales detail to ensure the sale was recorded in the proper period.**

Perform Test of Details – Traditional vs. New approach



Sampling

Processing Few Notable Items

Full Population

Processing Numerous Notable Items

Multidimensional Audit Data Selection - MADS

- **Outlier Detection Technique** – Use risk criteria (buckets) to prioritize

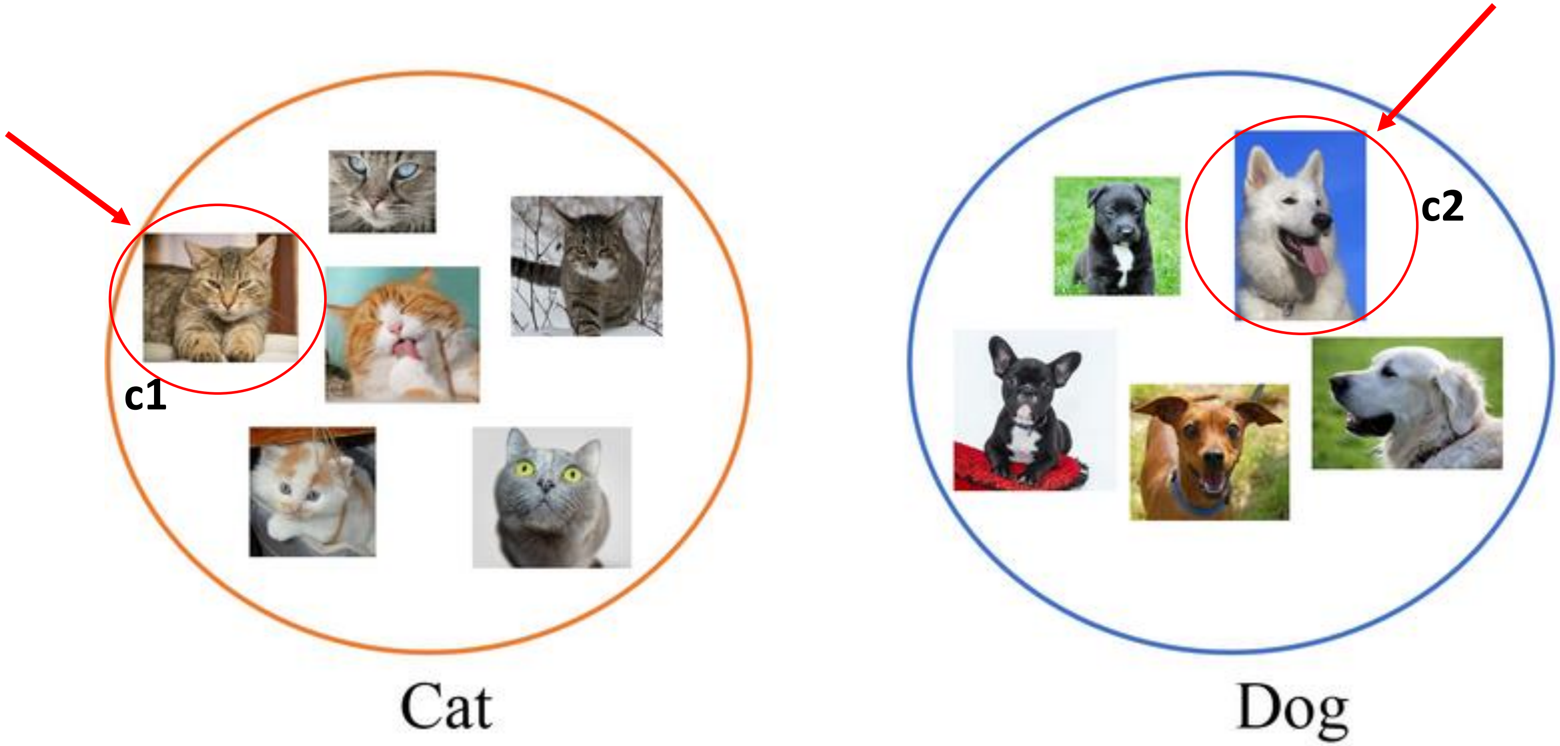
Type of Risk	Risk Level – Qualitative	Risk Level – Quantitative
Price Difference	High	100%
Date Difference	Medium	67%
Quantity Difference	Low	33%

Human involvement - Determined by auditor

Next step – program these set of inputs and apply them to each transaction to come up with **TOTAL RISK SCORE**

Clustering for Outlier Detection

- Clustering analysis is a data mining methodology
- Groups sets of objects together into “clusters”
  - Minimizing the within group differences
  - Maximizing the inter-group differences



Clustering Using K – Medians Algorithm

- The K-medians algorithm operates on a set (X) of n points.
  - There were 11 photos of animals in prior slide
- It chooses k centers {c1, c2, …,ck } from X
  - 2 centers were chosen at random c1 and c2 from the 11 photos
- And forms k clusters {C1, C2, …, Ck}
  - 2 clusters were formed C1 = Cat; C2 = Dog by grouping the remaining photos based on similarity in characteristics (nose, mouth, ears) to the chosen centers
- It minimizes the sum of the distances from each xt to the center of its clusters ck.
  - Minimize the difference (nose, mouth, and ear size) between each animal photo and the center photos for each of the 2 clusters

Walmart Case Example

Clustering for Outlier Detection

- Total quantitative and qualitative exceptions for revenue tests = **345** observations

Part 1:

- Cluster full sample of exceptions based on the following ***quantitative*** characteristics:
    - DIF_AMT
    - DIF_QUANTITY
    - DIF_PRICE
    - SHIP_QUANTITY
    - SHIP_UNIT_COST

Clustering for Outlier Detection

- Invoice amount differences = **250** observations

Part 2:

- Cluster invoice amount differences based on the following ***quantitative*** characteristics:
    - DIF_AMT
    - DIF_QUANTITY
    - DIF_PRICE
    - SHIP_QUANTITY
    - SHIP_UNIT_COST

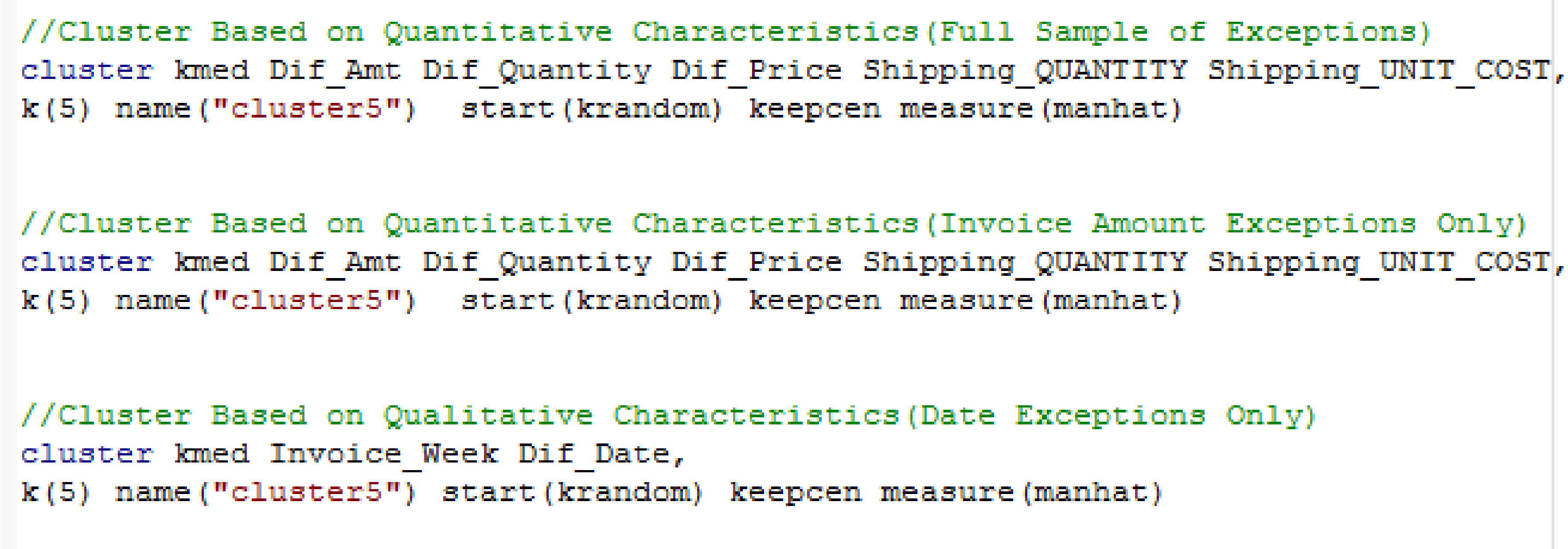
Clustering for Outlier Detection

- Date differences = **100** observations

Part 3:

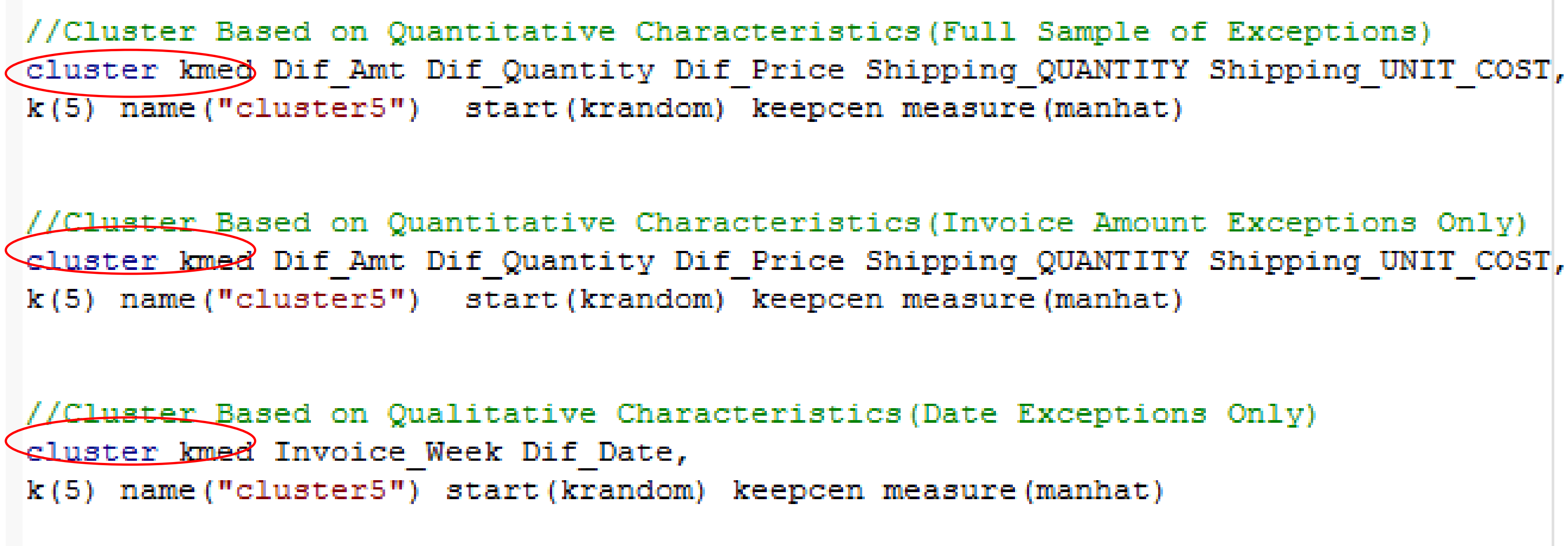
- Cluster date differences based on the following ***qualitative*** characteristics:
    - INVOICE_WEEK
    - DIF_DATE

Programming in Stata



```
//Cluster Based on Quantitative Characteristics(Full Sample of Exceptions)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Quantitative Characteristics(Invoice Amount Exceptions Only)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Qualitative Characteristics(Date Exceptions Only)
cluster kmed Invoice_Week Dif_Date,
k(5) name("cluster5") start(krandom) keepcen measure(manhat)
```

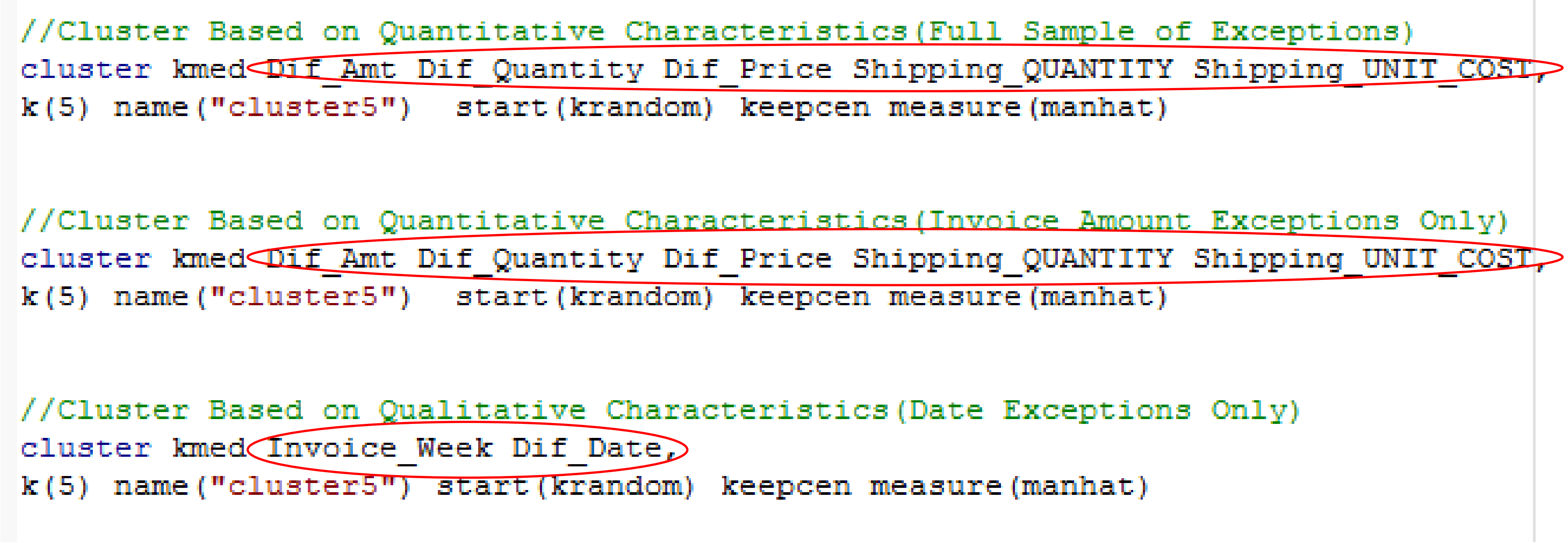
Programming in Stata



```
//Cluster Based on Quantitative Characteristics(Full Sample of Exceptions)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Quantitative Characteristics(Invoice Amount Exceptions Only)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Qualitative Characteristics(Date Exceptions Only)
cluster kmed Invoice_Week Dif_Date,
k(5) name("cluster5") start(krandom) keepcen measure(manhat)
```

- **cluster kmed** performs kmedians partition cluster analysis.

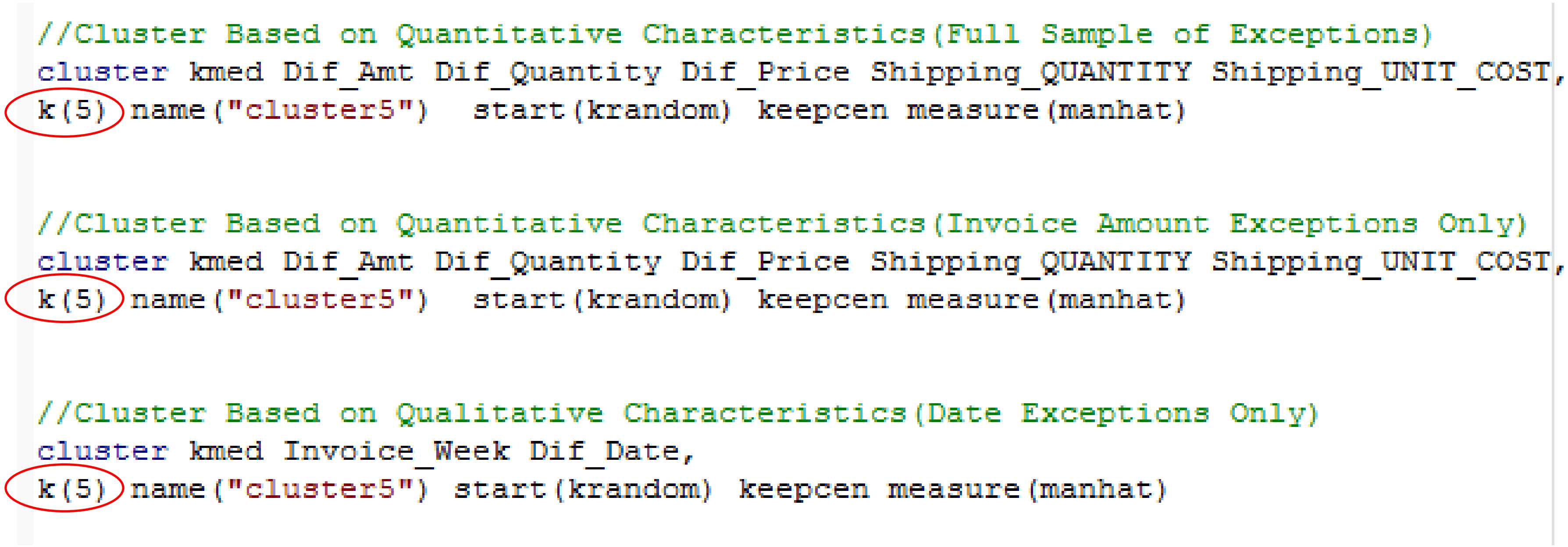
Programming in Stata



```
//Cluster Based on Quantitative Characteristics(Full Sample of Exceptions)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Quantitative Characteristics(Invoice Amount Exceptions Only)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Qualitative Characteristics(Date Exceptions Only)
cluster kmed Invoice_Week Dif_Date,
k(5) name("cluster5") start(krandom) keepcen measure(manhat)
```

- List of characteristics chosen to form clusters

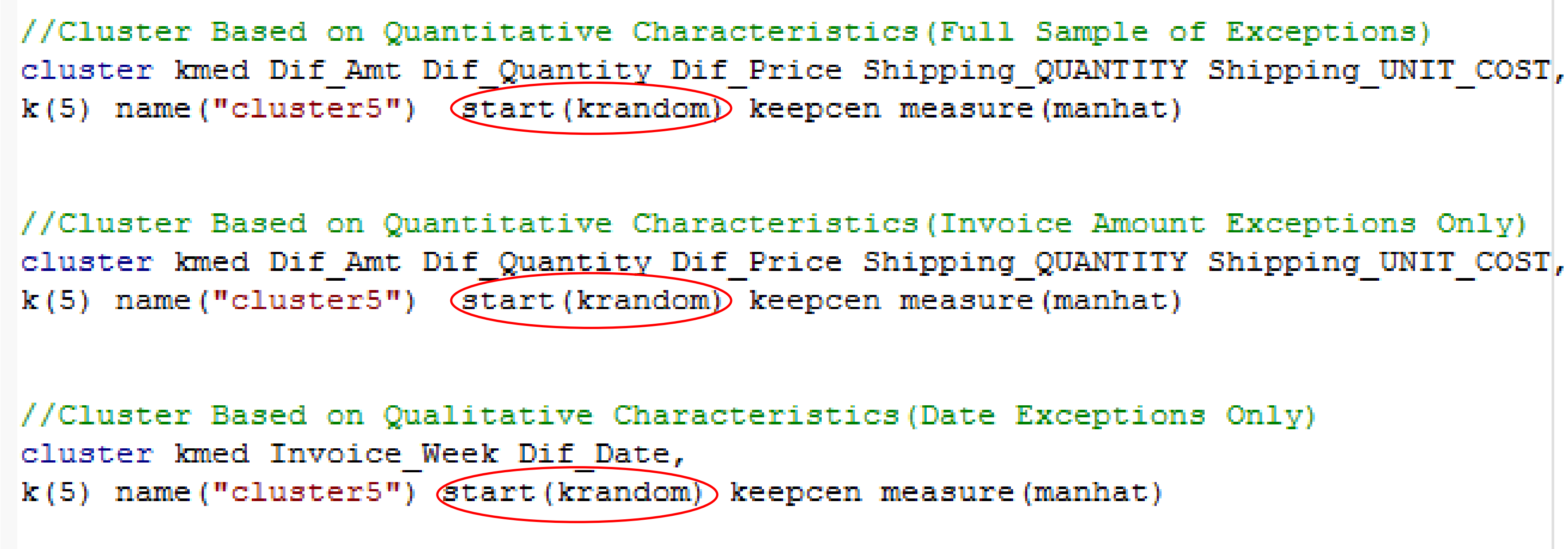
Programming in Stata



```
//Cluster Based on Quantitative Characteristics(Full Sample of Exceptions)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)

//Cluster Based on Quantitative Characteristics(Invoice Amount Exceptions Only)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)

//Cluster Based on Qualitative Characteristics(Date Exceptions Only)
cluster kmed Invoice_Week Dif_Date,
k(5) name("cluster5") start(krandom) keepcen measure(manhat)
```

- **k(#)** is required and indicates that # groups are to be formed by the cluster analysis.

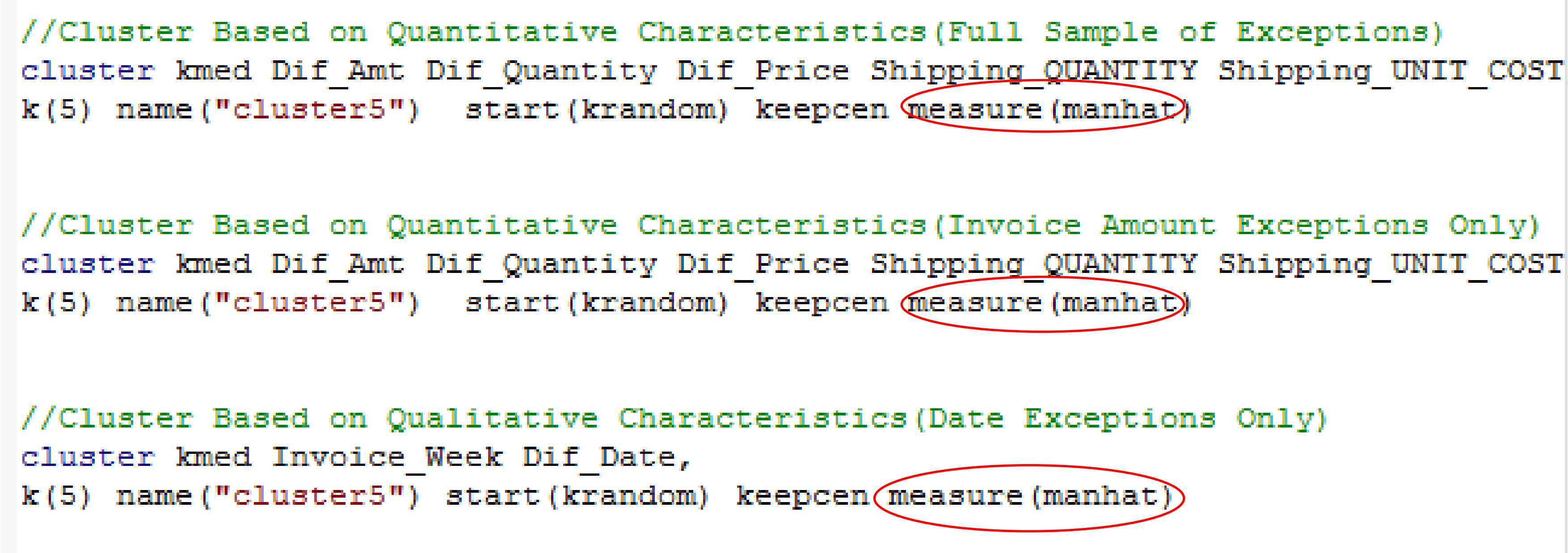
Programming in Stata



```
//Cluster Based on Quantitative Characteristics(Full Sample of Exceptions)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Quantitative Characteristics(Invoice Amount Exceptions Only)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Qualitative Characteristics(Date Exceptions Only)
cluster kmed Invoice_Week Dif_Date,
k(5) name("cluster5") start(krandom) keepcen measure(manhat)
```

- **start(krandom)** obtain k initial group centers chosen at random from the sample of observations

Programming in Stata

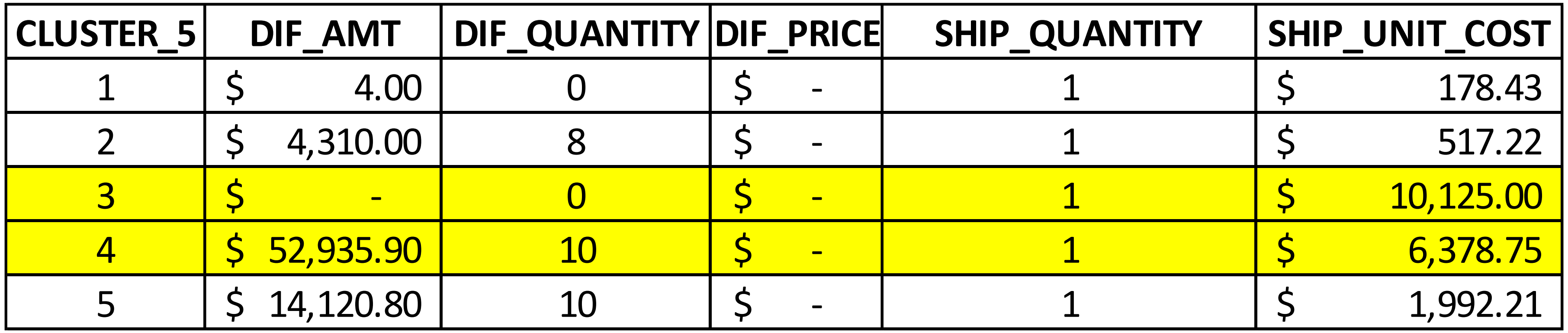


```
//Cluster Based on Quantitative Characteristics(Full Sample of Exceptions)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Quantitative Characteristics(Invoice Amount Exceptions Only)
cluster kmed Dif_Amt Dif_Quantity Dif_Price Shipping_QUANTITY Shipping_UNIT_COST,
k(5) name("cluster5")  start(krandom) keepcen measure(manhat)


//Cluster Based on Qualitative Characteristics(Date Exceptions Only)
cluster kmed Invoice_Week Dif_Date,
k(5) name("cluster5") start(krandom) keepcen measure(manhat)
```

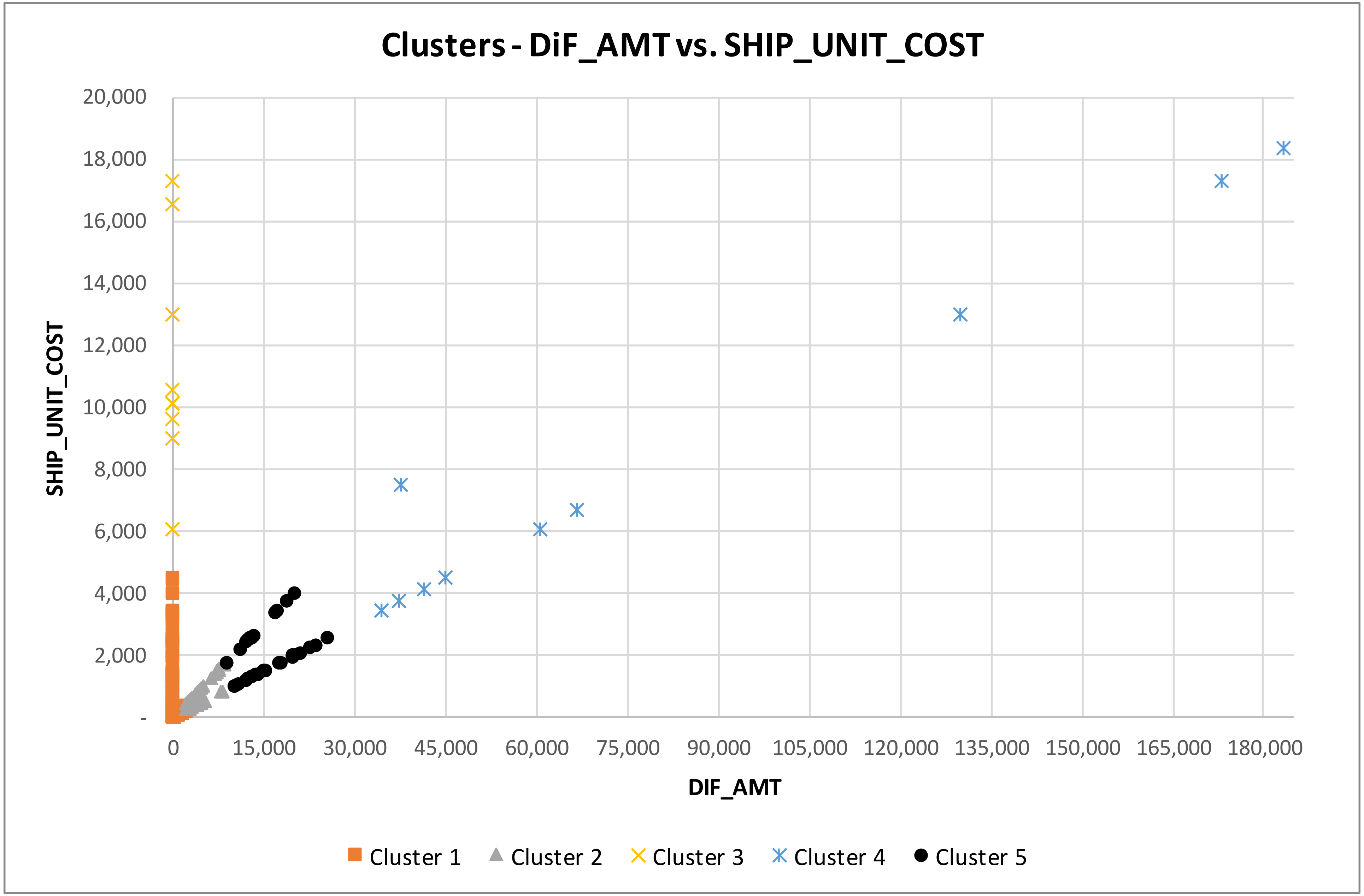
- **measure(manhat)** specifies the similarity or dissimilarity measure. Here, we use Manhattan distance. Our goal is to minimize the distance between the datapoints and their cluster centers.

Part 1: Cluster Full Sample of Exceptions

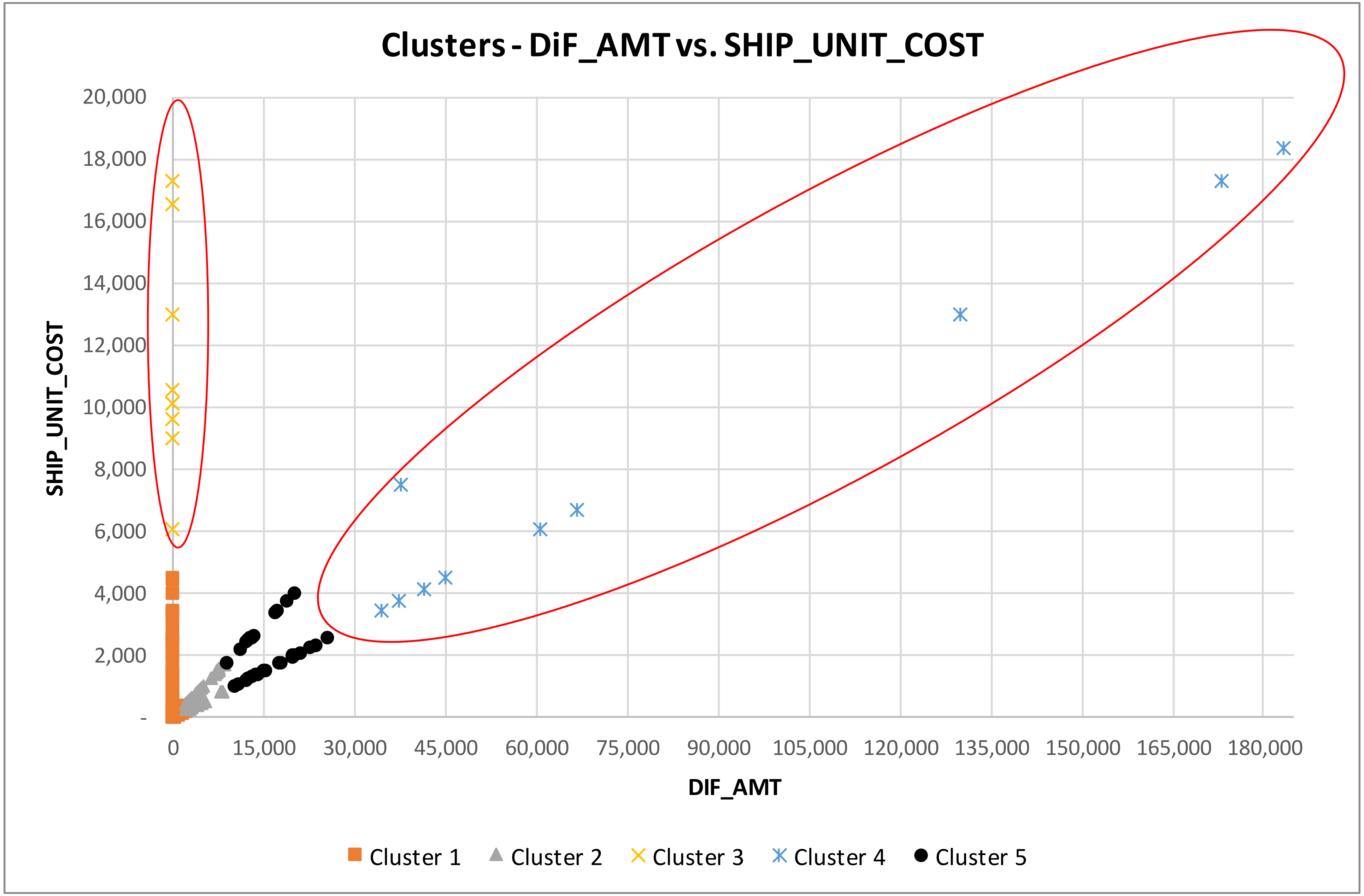
- Cluster Medians



CLUSTER_5	DIF_AMT	DIF_QUANTITY	DIF_PRICE	SHIP_QUANTITY	SHIP_UNIT_COST
1	$ 4.00	0	$ -	1	$ 178.43
2	$ 4,310.00	8	$ -	1	$ 517.22
3	$ -	0	$ -	1	$ 10,125.00
4	$ 52,935.90	10	$ -	1	$ 6,378.75
5	$ 14,120.80	10	$ -	1	$ 1,992.21

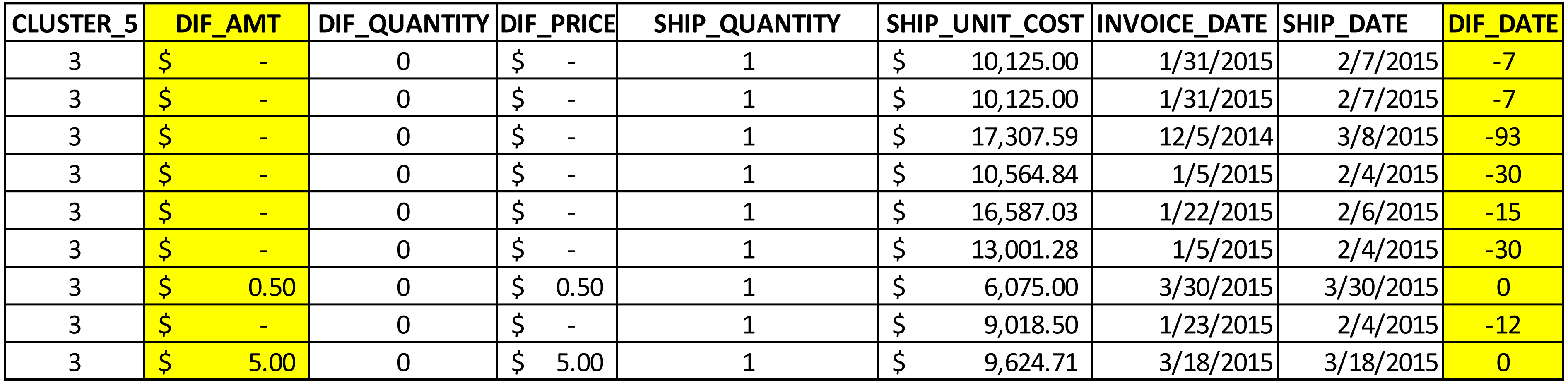


Clusters - DiF_AMT vs. SHIP_UNIT_COST


Clusters - DiF_AMT vs. SHIP_UNIT_COST
SHIP_UNIT_COST
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
-
0
15,000
30,000
45,000
60,000
75,000
90,000
105,000
120,000
135,000
150,000
165,000
180,000
DIF_AMT
■ Cluster 1
▲ Cluster 2
× Cluster 3
✳ Cluster 4
● Cluster 5

Part 1: Cluster Full Sample of Exceptions

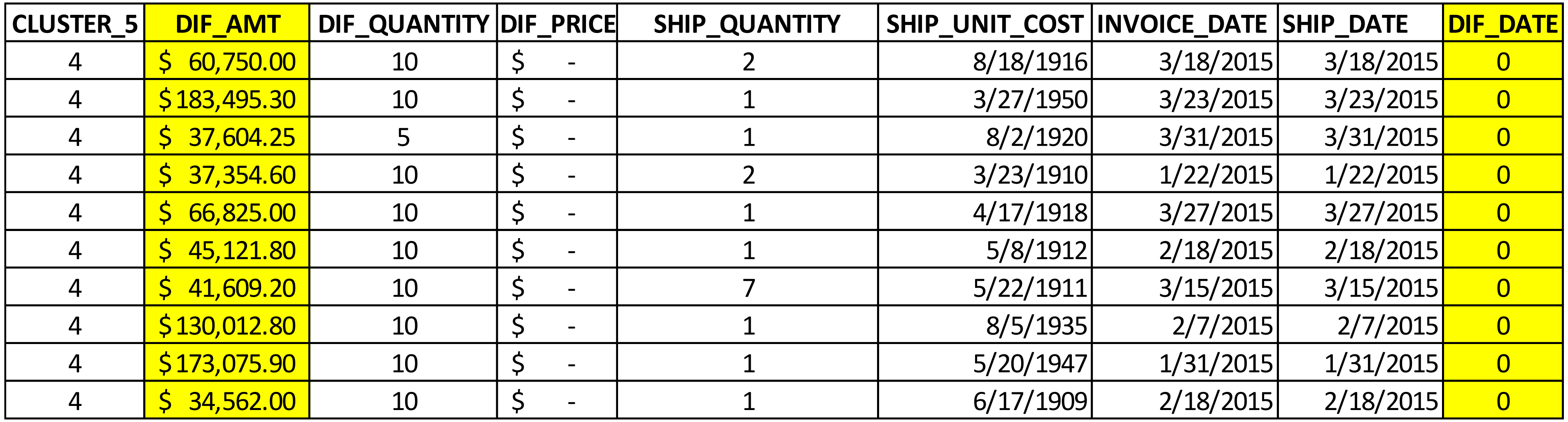
- Cluster 3



CLUSTER_5	DIF_AMT	DIF_QUANTITY	DIF_PRICE	SHIP_QUANTITY	SHIP_UNIT_COST	INVOICE_DATE	SHIP_DATE	DIF_DATE
3	$ -	0	$ -	1	$ 10,125.00	1/31/2015	2/7/2015	-7
3	$ -	0	$ -	1	$ 10,125.00	1/31/2015	2/7/2015	-7
3	$ -	0	$ -	1	$ 17,307.59	12/5/2014	3/8/2015	-93
3	$ -	0	$ -	1	$ 10,564.84	1/5/2015	2/4/2015	-30
3	$ -	0	$ -	1	$ 16,587.03	1/22/2015	2/6/2015	-15
3	$ -	0	$ -	1	$ 13,001.28	1/5/2015	2/4/2015	-30
3	$ 0.50	0	$ 0.50	1	$ 6,075.00	3/30/2015	3/30/2015	0
3	$ -	0	$ -	1	$ 9,018.50	1/23/2015	2/4/2015	-12
3	$ 5.00	0	$ 5.00	1	$ 9,624.71	3/18/2015	3/18/2015	0

Part 1: Cluster Full Sample of Exceptions

- Cluster 4



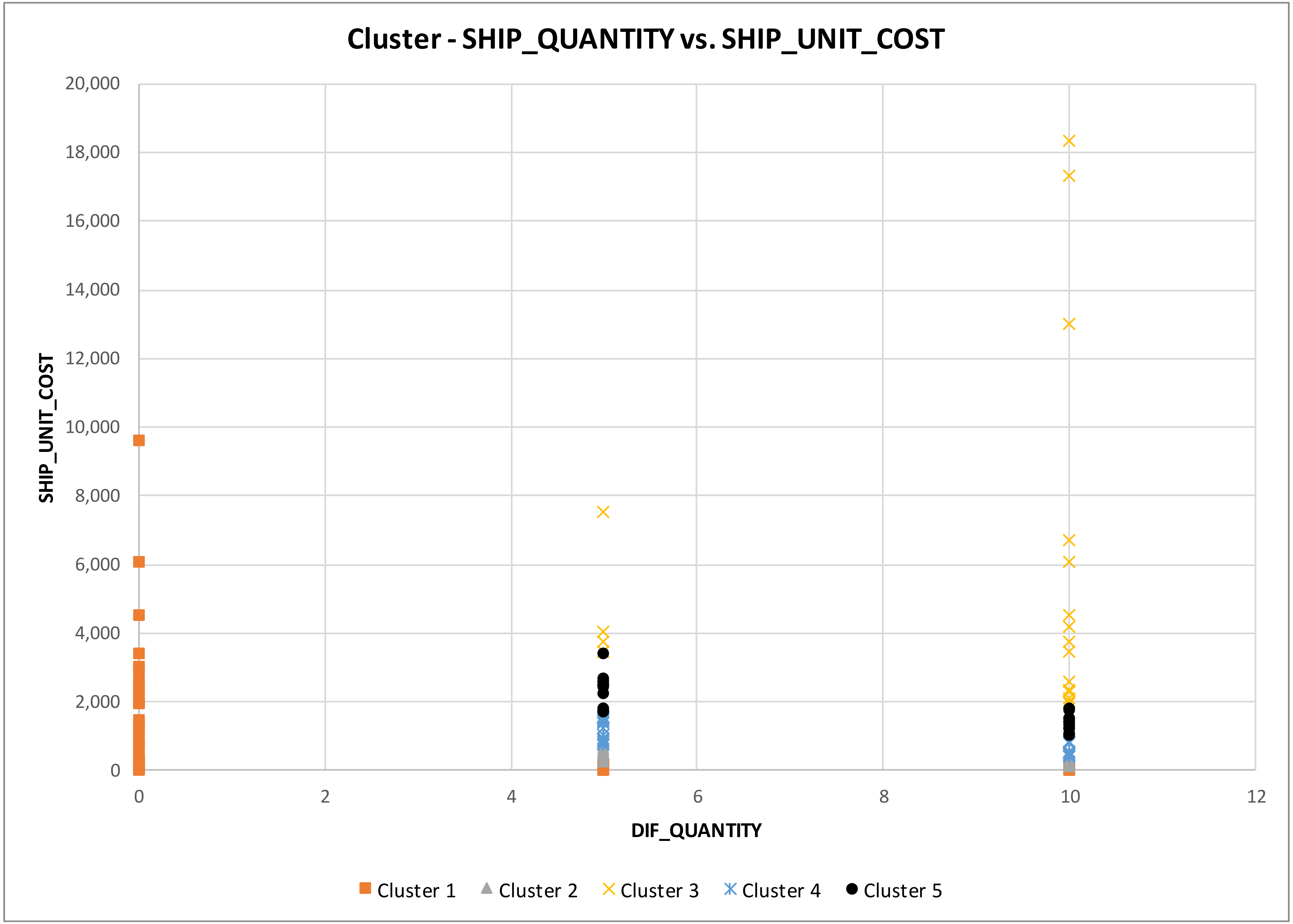
CLUSTER_5	DIF_AMT	DIF_QUANTITY	DIF_PRICE	SHIP_QUANTITY	SHIP_UNIT_COST	INVOICE_DATE	SHIP_DATE	DIF_DATE
4	$ 60,750.00	10	$ -	2	8/18/1916	3/18/2015	3/18/2015	0
4	$ 183,495.30	10	$ -	1	3/27/1950	3/23/2015	3/23/2015	0
4	$ 37,604.25	5	$ -	1	8/2/1920	3/31/2015	3/31/2015	0
4	$ 37,354.60	10	$ -	2	3/23/1910	1/22/2015	1/22/2015	0
4	$ 66,825.00	10	$ -	1	4/17/1918	3/27/2015	3/27/2015	0
4	$ 45,121.80	10	$ -	1	5/8/1912	2/18/2015	2/18/2015	0
4	$ 41,609.20	10	$ -	7	5/22/1911	3/15/2015	3/15/2015	0
4	$ 130,012.80	10	$ -	1	8/5/1935	2/7/2015	2/7/2015	0
4	$ 173,075.90	10	$ -	1	5/20/1947	1/31/2015	1/31/2015	0
4	$ 34,562.00	10	$ -	1	6/17/1909	2/18/2015	2/18/2015	0

Part 2: Cluster Invoice Amount Differences Only

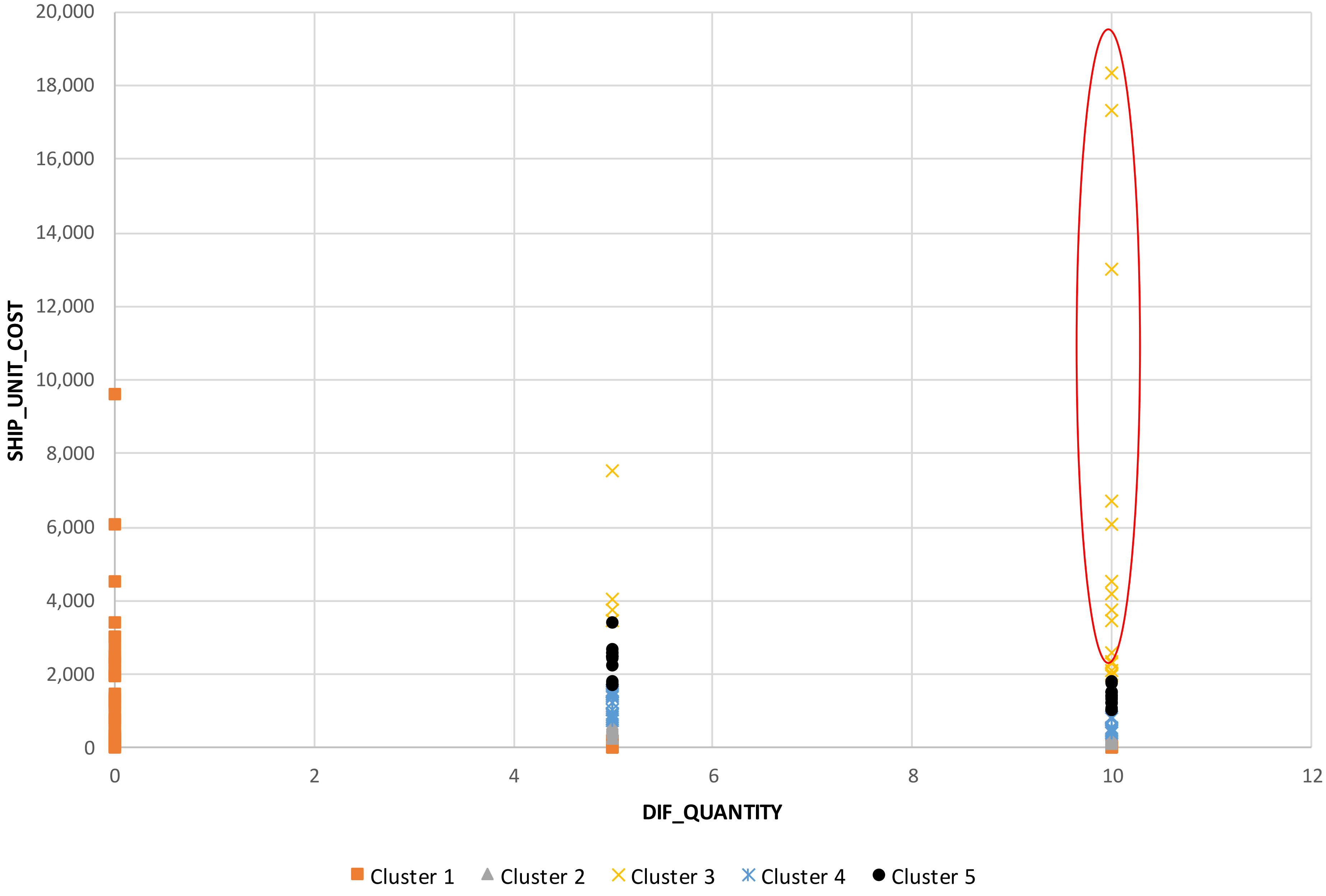
- Cluster Medians

CLUSTER_5	DIF_AMT	DIF_QUANTITY	DIF_PRICE	SHIP_QUANTITY	SHIP_UNIT_COST
1	$ 32.00	5	$ -	1	$ 71.00
2	$ 2,210.18	10	$ -	1	$ 289.32
3	$ 23,455.40	10	$ -	1	$ 3,456.20
4	$ 5,172.20	5	$ -	1	$ 821.99
5	$ 12,502.55	10	$ -	1	$ 1,614.34

High Risk	High unit cost and high quantity difference
Moderate Risk	High unit cost and low quantity difference
Moderate Risk	Low unit cost and high quantity difference
Low Risk	Low unit cost and low quantity difference


Cluster - SHIP_QUANTITY vs. SHIP_UNIT_COST
SHIP_UNIT_COST
20,000
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
0
2
4
6
8
10
12
DIF_QUANTITY
■ Cluster 1
▲ Cluster 2
× Cluster 3
× Cluster 4
● Cluster 5

Cluster - SHIP_QUANTITY vs. SHIP_UNIT_COST



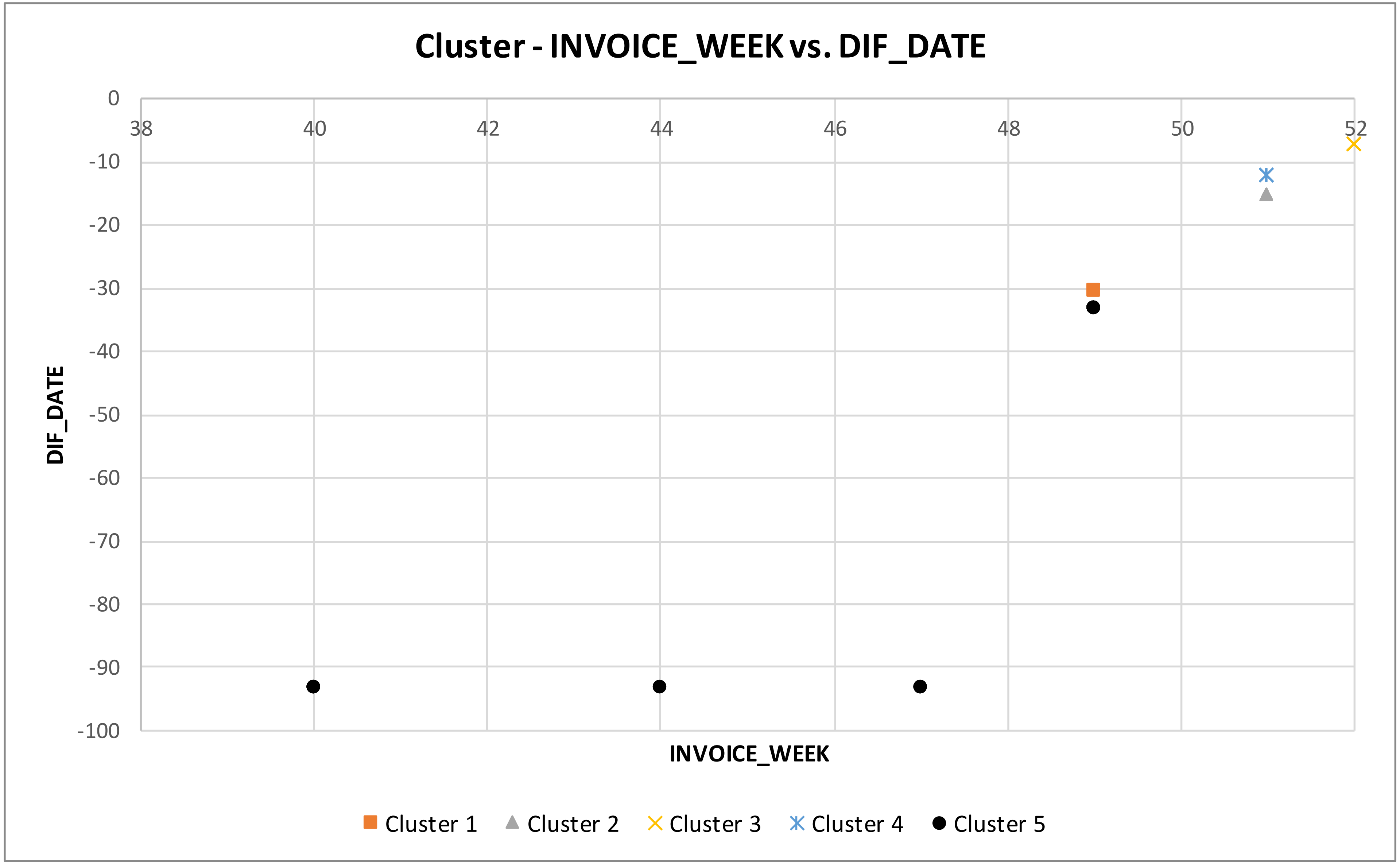
CLUSTER_5	DIF_AMT	DIF_QUANTITY	DIF_PRICE	SHIP_QUANTITY	SHIP_UNIT_COST
3	$ 19,712.60	10	$ -	1	$ 1,971.26
3	$ 19,756.50	10	$ -	1	$ 1,975.65
3	$ 25,643.40	10	$ -	1	$ 2,564.34
3	$ 20,140.25	5	$ -	1	$ 4,028.05
3	$ 60,750.00	10	$ -	2	$ 6,075.00
3	$ 183,495.30	10	$ -	1	$ 18,349.53
3	$ 19,922.10	10	$ -	1	$ 1,992.21
3	$ 23,455.40	10	$ -	1	$ 2,345.54
3	$ 17,281.00	5	$ -	1	$ 3,456.20
3	$ 37,604.25	5	$ -	1	$ 7,520.85
3	$ 37,354.60	10	$ -	2	$ 3,735.46
3	$ 66,825.00	10	$ -	1	$ 6,682.50
3	$ 18,832.90	5	$ -	1	$ 3,766.58
3	$ 45,121.80	10	$ -	1	$ 4,512.18
3	$ 22,788.70	10	$ -	3	$ 2,278.87
3	$ 21,000.00	10	$ -	1	$ 2,100.00
3	$ 19,922.10	10	$ -	1	$ 1,992.21
3	$ 41,609.20	10	$ -	7	$ 4,160.92
3	$ 130,012.80	10	$ -	1	$ 13,001.28
3	$ 17,281.00	5	$ -	6	$ 3,456.20
3	$ 173,075.90	10	$ -	1	$ 17,307.59
3	$ 34,562.00	10	$ -	1	$ 3,456.20
3	$ 23,455.40	10	$ -	1	$ 2,345.54
3	$ 20,020.30	10	$ -	1	$ 2,002.03
Total Amt Dif	**$ 1,099,622.50**				

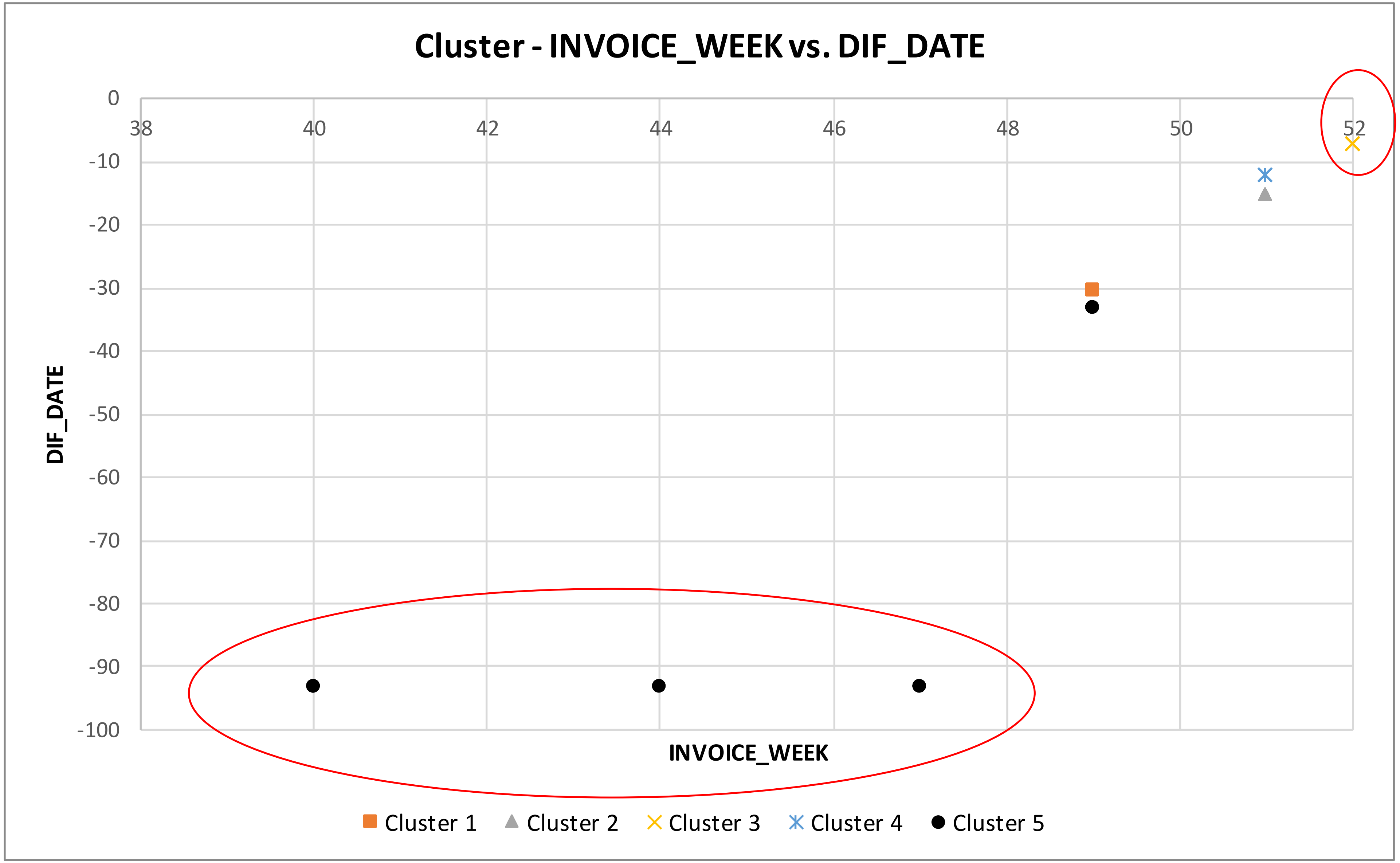
Part 3: Cluster Date Differences Only

- Cluster Medians

CLUSTER_5	INVOICE_WEEK	DIF_DATE
1	49	-30
2	51	-15
3	52	-7
4	51	-12
5	49	-33

High Risk Small date difference in last week of the year - RMM due to fraud
High Risk Large date difference - RMM due to systematic errors
Moderate Risk Date difference in second to last week of the year - RMM due to fraud or systematic errors


Cluster - INVOICE_WEEK vs. DIF_DATE
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
38
40
42
44
46
48
50
52
DIF_DATE
INVOICE_WEEK
■ Cluster 1
▲ Cluster 2
✕ Cluster 3
✳ Cluster 4
● Cluster 5


Cluster - INVOICE_WEEK vs. DIF_DATE
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
38
40
42
44
46
48
50
52
DIF_DATE
INVOICE_WEEK
■ Cluster 1
▲ Cluster 2
✕ Cluster 3
✳ Cluster 4
● Cluster 5

CLUSTER_5	INVOICE_WEEK	DIF_DATE	INVOICE_GROSS_AMT	INVOICE_DATE	SHIP_DATE
3	52	-7	$ 283.98	1/31/2015	2/7/2015
3	52	-7	$ 14,422.32	1/31/2015	2/7/2015
3	52	-7	$ 286.14	1/31/2015	2/7/2015
3	52	-7	$ 431.80	1/31/2015	2/7/2015
3	52	-7	$ 10,125.00	1/31/2015	2/7/2015
3	52	-7	$ 10,125.00	1/31/2015	2/7/2015
3	52	-7	$ 0.01	1/31/2015	2/7/2015
3	52	-7	$ 2,447.82	1/31/2015	2/7/2015
3	52	-7	$ 7.29	1/31/2015	2/7/2015
3	52	-7	$ 29.78	1/29/2015	2/5/2015
3	52	-7	$ 3,851.60	1/31/2015	2/7/2015
3	52	-7	$ 13.85	1/31/2015	2/7/2015
3	52	-7	$ 98.14	1/29/2015	2/5/2015
3	52	-7	$ 10.11	1/31/2015	2/7/2015
3	52	-7	$ 592.83	1/31/2015	2/7/2015
3	52	-7	$ 173.59	1/29/2015	2/5/2015
3	52	-7	$ 41.33	1/31/2015	2/7/2015
3	52	-7	$ 2,602.38	1/29/2015	2/5/2015
3	52	-7	$ 747.26	1/29/2015	2/5/2015
3	52	-7	$ 1,369.12	1/29/2015	2/5/2015
3	52	-7	$ 651.58	1/31/2015	2/7/2015
3	52	-7	$ 322.80	1/31/2015	2/7/2015
3	52	-7	$ 13,824.80	1/31/2015	2/7/2015
3	52	-7	$ 18.37	1/29/2015	2/5/2015
3	52	-7	$ 52.05	1/31/2015	2/7/2015
3	52	-7	$ 10.34	1/31/2015	2/7/2015
3	52	-7	$ 289.32	1/31/2015	2/7/2015
			$ 62,828.61		

CLUSTER_5	INVOICE_WEEK	DIF_DATE	INVOICE_GROSS_AMT	INVOICE_DATE	SHIP_DATE
5	49	-33	$ 692.72	1/4/2015	2/6/2015
5	49	-33	$ 16,112.20	1/4/2015	2/6/2015
5	49	-33	$ 2,658.51	1/4/2015	2/6/2015
5	49	-33	$ 154.78	1/4/2015	2/6/2015
5	49	-33	$ 2,564.34	1/4/2015	2/6/2015
5	40	-93	$ 2,416.83	11/4/2014	2/5/2015
5	49	-33	$ 95.95	1/4/2015	2/6/2015
5	44	-93	$ 17,307.59	12/5/2014	3/8/2015
5	40	-93	$ 1,026.61	11/4/2014	2/5/2015
5	49	-33	$ 0.11	1/4/2015	2/6/2015
5	40	-93	$ 234.51	11/4/2014	2/5/2015
5	49	-33	$ 127.80	1/4/2015	2/6/2015
5	49	-33	$ 455.78	1/4/2015	2/6/2015
5	44	-93	$ 21.00	12/5/2014	3/8/2015
5	40	-93	$ 2,829.82	11/4/2014	2/5/2015
5	49	-33	$ 1,245.43	1/4/2015	2/6/2015
5	49	-33	$ 57.86	1/4/2015	2/6/2015
5	47	-93	$ 7.18	12/22/2014	3/25/2015
5	40	-93	$ 7,472.58	11/4/2014	2/5/2015
5	49	-33	$ 4,028.05	1/4/2015	2/6/2015
5	49	-33	$ 1,702.58	1/4/2015	2/6/2015
5	49	-33	$ 1,001.67	1/4/2015	2/6/2015
5	47	-93	$ 455.78	12/22/2014	3/25/2015
5	49	-33	$ 2,500.51	1/4/2015	2/6/2015
			$ 18,471.64		

Summary and Takeaways

- Clustering can be used for ***outlier detection*** as part of substantive testing.

- Clustering can be used to categorize/rank/prioritize ***exceptions***.

- By grouping the data based on similarities in characteristics, auditors can utilize a ***targeted approach*** to address the ***specific risks*** related to each cluster.

- Clustering is a data driven technique that can help remove auditor bias.