Synthetic Data Generation In R: Pistil Storage Case Study

1

Professor Thomas McKee

Medical University of South Carolina

Norwegian School of Economics

temckee@musc.edu

Motivation

- Professional and academic organizations have embraced the importance of working with data as a core competency in the accounting/auditing profession. This focus has increased the hands-on use of audit data analytics in teaching auditing and generated a demand for realistic data to use in the analytics projects.

- Real data is not always easy to obtain and may contain problems such as confidentiality agreements, data cleaning, and lack of desired data frequencies (e.g., not enough fraud examples).

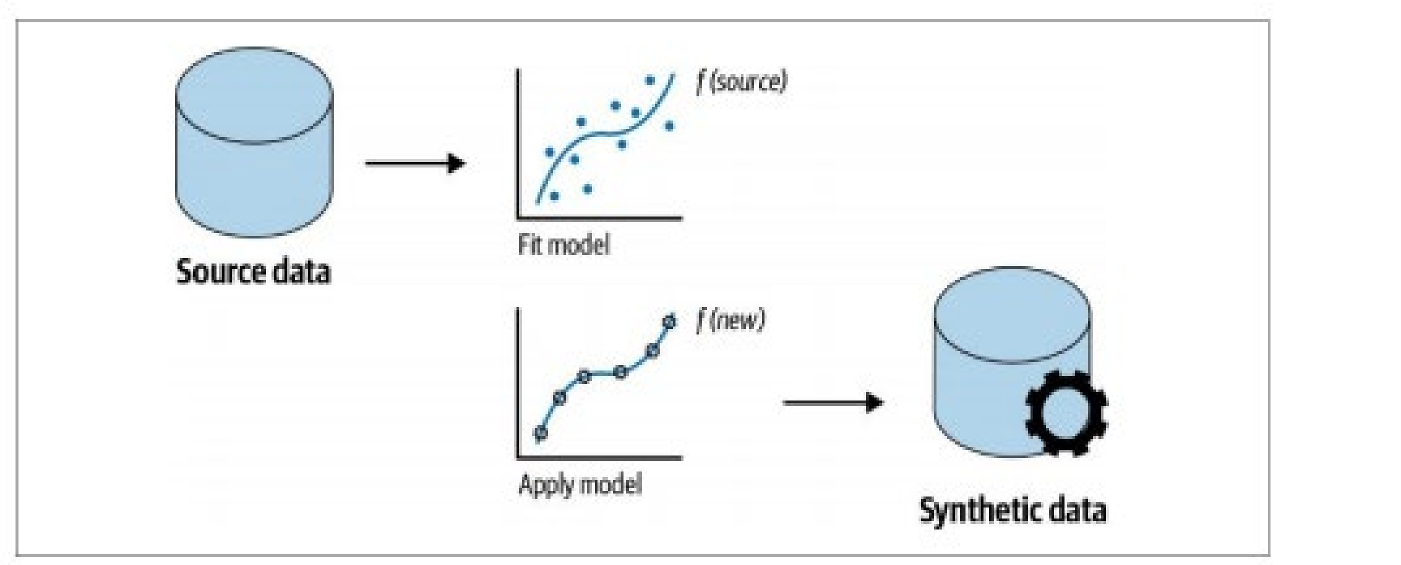
Synthetic Data Advantages

- Synthetic cata can be used to create replacement data sets for confidential data
- Fraud or anomalies can be inserted into the data in predetermined amounts or frequencies
- High-dimensional data [extremely large] can be simulated when volume of real data is limited
- Data can be created with relatively little code.
- Data bias exists in many real datasets
- Real data may lack data elements which exist in other data sources outside the real data source.

Types of Synthetic Data

- **Synthesis from Real Data -**uses distributions and structure of real data. This is illustrated with the figure on the following slide.
    - *Fully Synthetic*-does not contain any real data
    - *Partially Synthetic*-noise/data is induced into real data to simulate additional cases
- **Synthesis Without Real Data** -uses existing models or analyst's background knowledge
- **Hybrid Approach** -uses simplifying assumptions about real data

Synthetic Data From Real Data



Why R ?

- R is a well-developed, simple, widely used open source programming language for statistical computing and graphical techniques.

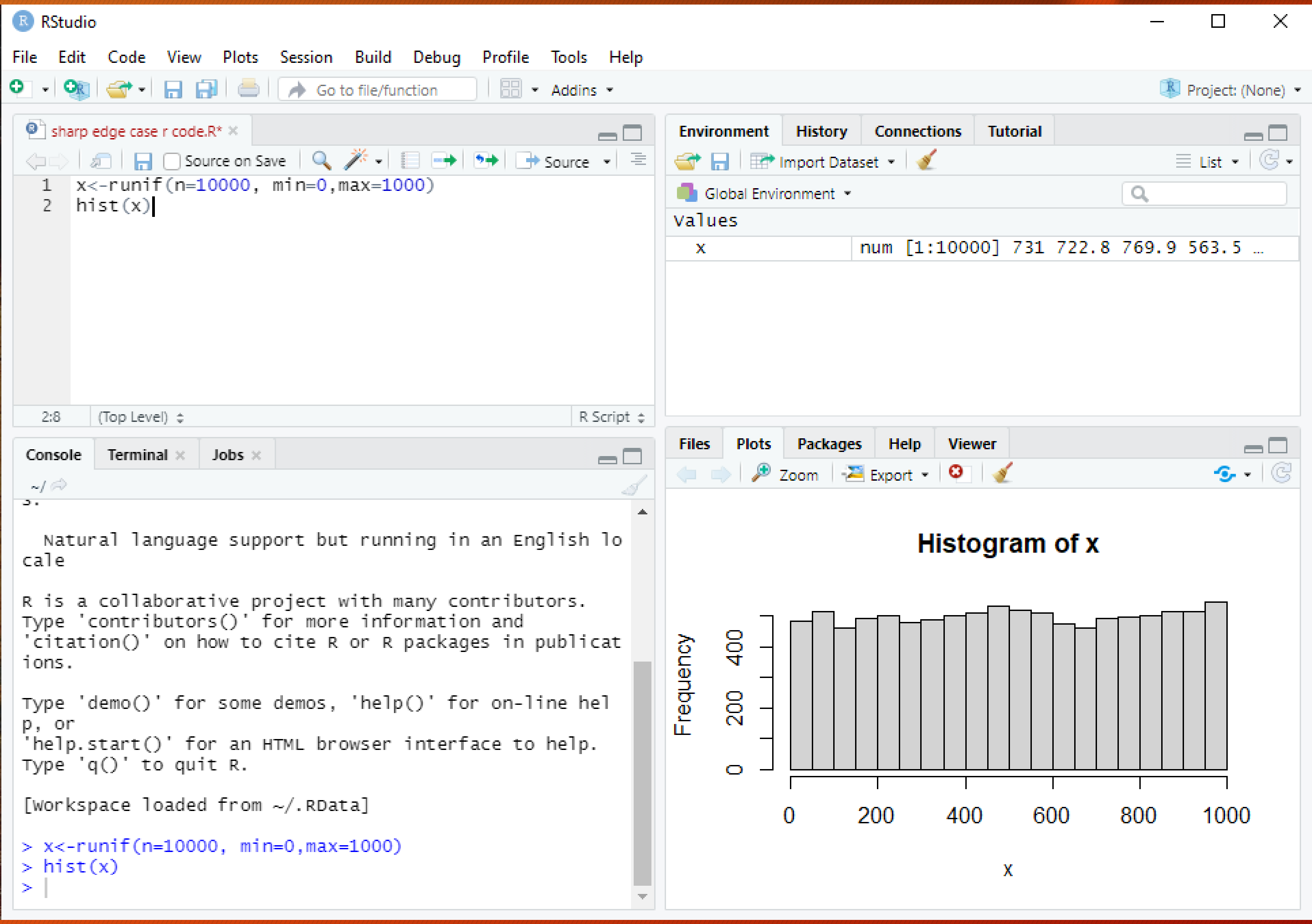
- R runs in the free RStudio integrated development environment as well as other software platforms.

- R is an interpreted language which can execute written code on a line by line basis. This can be a big advantage in data creation as it enables output to be examined immediately to see if it is the desired output.

- RStudio includes a user-friendly desktop console, syntax highlighting editor as well as plotting, debugging, and workspace management tools.

RStudio Console Example-

Create 10,000 random numbers between 0 -1,000 and create histogram of results



```
x<-runif(n=10000, min=0,max=1000)
hist(x)
```

```
Values
  x                 num [1:10000] 731 722.8 769.9 563.5 ...
```

```
3.

  Natural language support but running in an English lo
cale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publicat
ions.

Type 'demo()' for some demos, 'help()' for on-line hel
p, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> x<-runif(n=10000, min=0,max=1000)
> hist(x)
> 
```

R Simulation Packages

A wide variety of previously developed simulation packages which automate aspects of data simulation are available:

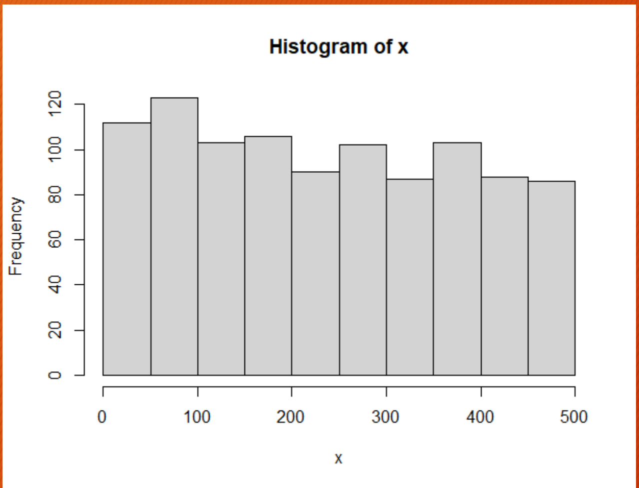
- simFrame
- simPop
- Tidyverse
- Bindata
- Charlatan
- fakeR
- PoisBinOrdNonNor
- SimMultiCorrData

R Data Distributions

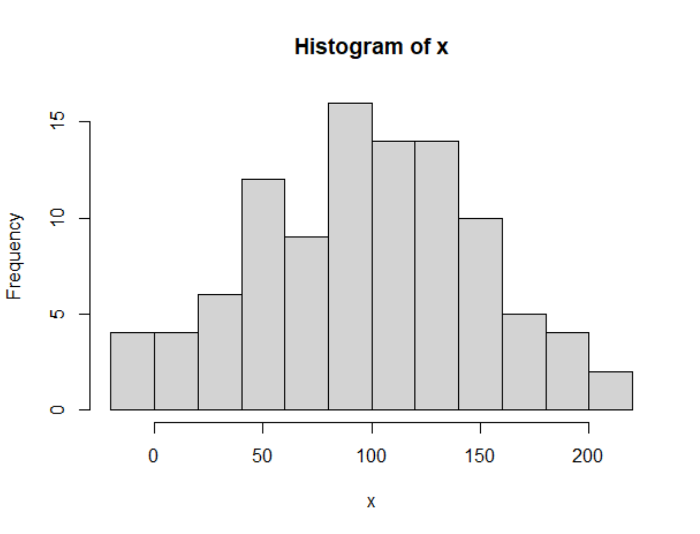
- R has the ability to create almost any data distributions
- You are primarily only limited by your R programming knowledge
- Common distributions in R
  - Binomial distribution
  - Normal distribution
  - t distribution
  - Chi squared
  - F
  - Poisson distribution
  - Uniform
-

Uniform and Normal Data Distribution Code and Output Graphs

- Uniform
  - “ x < - runif (n=1000, min=0, max =500)”

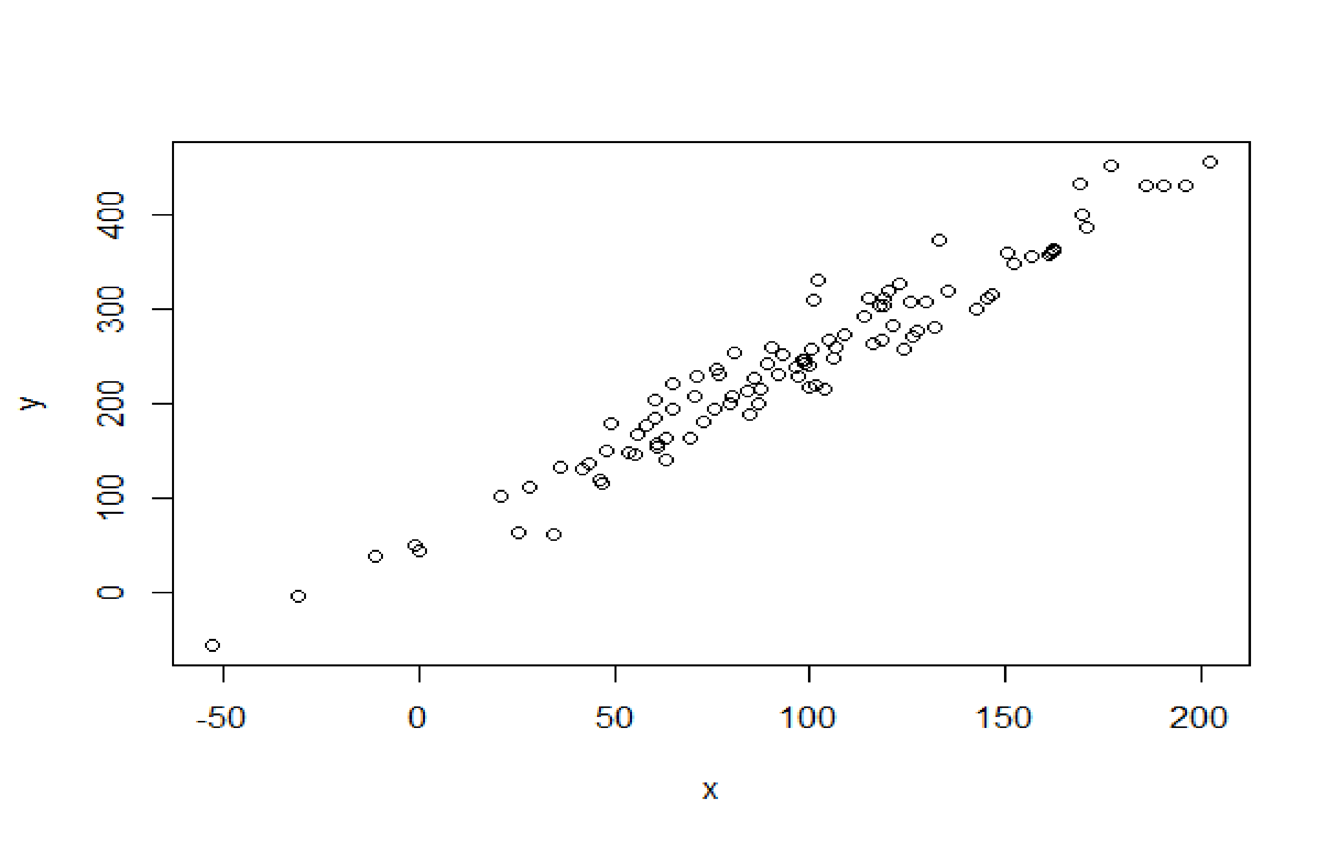


- Normal
  - “x <-rnorm (n=100, mean=100, sd=50)”



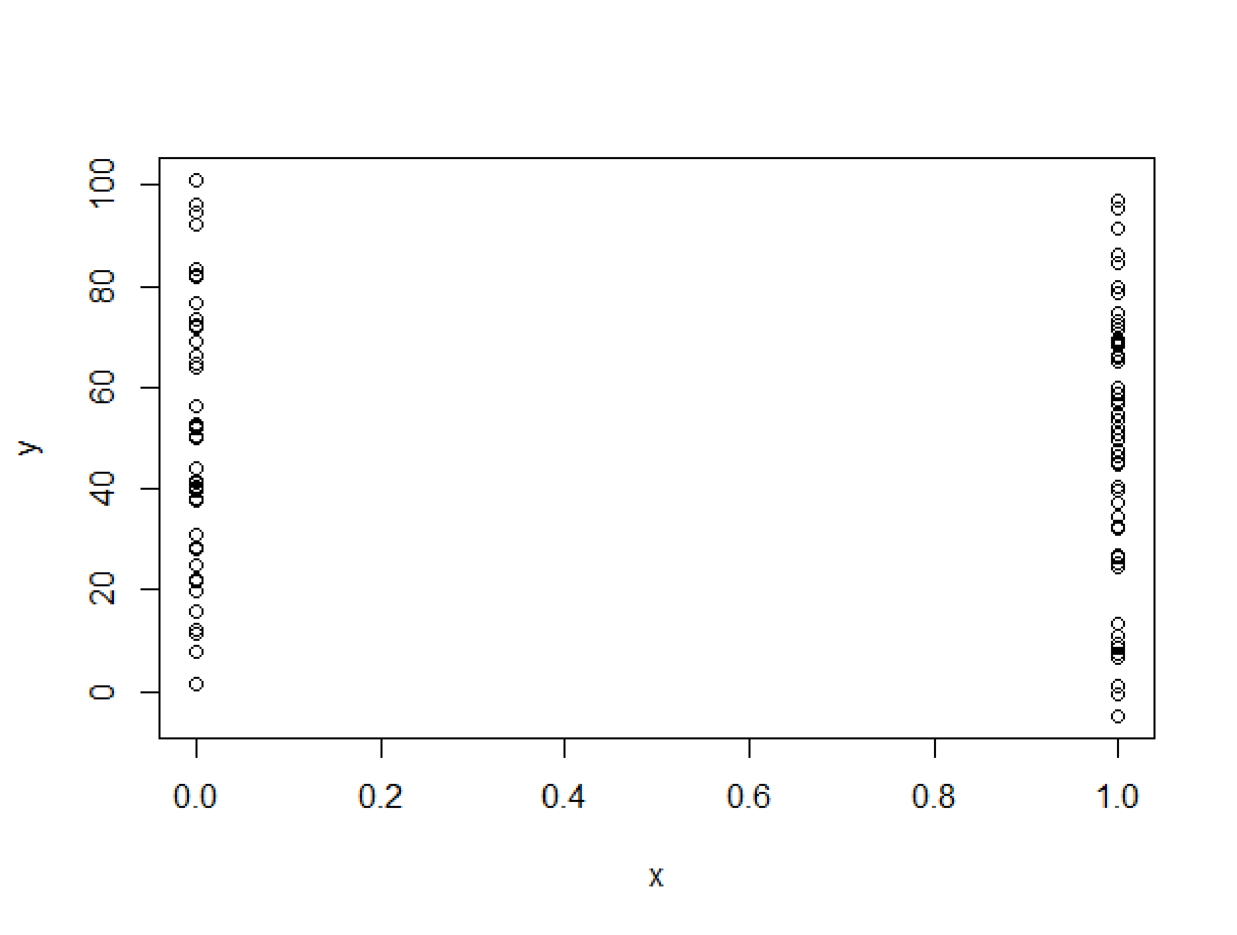
Simple Linear Model

- x <- rnorm(n=100, mean=100, sd=50)
- e <-rnorm (100, mean =50, sd=25)
- y <- .05 + (2 * x) + e



Binary Data

- x<- rbinom (100, 1, .05)
- str(x)
- e <-rnorm (100, mean =50, sd=25)
- y <- .05 + (2 * x) + e



Steps In Creation of Pistil Storage Case

1. Decision on case learning objectives and type of analysis to be undertaken, e.g. cluster analysis

2. Searching SEC filings to find background data for a suitable company

3. Write the case commentary using company background but changing company name and data

4. Creation of data
    1. Generate 5 sets of revenue data for 5 company divisions with as many related variables as needed in each set
    2. Seed revenue errors into one division.
    3. Seed payroll errors into second division
    4. Convert each data set to dataframe via single line dataframe command
    5. Combine the 5 divisional dataframes into a single data dataframe for the students via Single R command

        - Pistil <-rbind(dataframe1,dataframe2,dataframe3, dataframe4, dataframe5)

    6. Output data as CSV file for easy student access

5. Computing case solutions

6. Repeat steps 4 and 5 until satisfied results provide reasonable case

Pistil Storage Brief Overview

- Pistil Storage Inc. is a publicly traded audit client listed on the NYSE.
- Two troubling phone calls were received during 2017 on the whistleblower hotline:
    - Length of employment per payroll records may have errors
    - Cash payments for first month's rent when rent is normally paid via credit/debit card.
- There is apparently no direct way to follow up on the specific call allegations.
- You have been selected to apply exploratory ADA via cluster analysis to revenue data to see if any anomalies are apparent.

Pistil Storage Case

- Pistil Storage Inc. is a fully integrated Delaware corporation which was formed on January 1,1999, to own, operate, manage, acquire, develop and redevelop professionally managed self-storage properties (“stores”).Common stock is traded on the New York Stock Exchange under the symbol “PIS.”
- “Self-storage” lets a person or business store things which the person or business does not have space for at a storage facility (“store”). Storage facilities rent storage space (typically called “storage units” or “storage lockers” to tenants on a short-term basis, sometimes month-to-month, but often on one year leases. Larger storage facilities may offer access 24-hours a day and 7 days a week. Storage facilities typically have a variety of security features to protect the storage units. Many facilities require that the tenant secure their unit with their own lock and key so only the tenant has actual access to the unit. Self-storage is a rapidly growing industry, as in addition to individuals storing things, many businesses rent storage units to store inventory or equipment.
- The self-storage industry is characterized by fragmented ownership. The top ten self-storage companies in the United States operated approximately 20.1% of the total U.S. stores, and the top 50 self-storage companies operated approximately 30.2% of the total U.S. stores as of December 31, 2017.
- Pistil Storage owned and operated 1531stores at December 31, 2017. These stores are located in 44 states and contain approximately 37 million square feet of net rentable space in approximately 306,000 storage units and currently serve a customer base of approximately 250,000 tenants. Recorded revenue for 2017 was $499,200,823.
- The company operates throughout the United States but divides the country into five distinct districts through a planned expansion plan which seeks to establish approximately the same number of stores in each district. Each of the five districts is managed by a district manager. The district managers receive a base salary plus a small bonus based on meeting or exceeding budget targets. The districts and number of stores in each district as of December 31, 2017 are shown in Exhibit 1:
- Exhibit 1-Store Number District Locations

District Store Number Range	District Name	Number of Stores In District
1-400	Southeast	306
401-800	Midwest	311
801-1200	Northeast	318
1201-1600	Southwest	299
1601-2000	West	297

Pistil Storage Case

- Many of the stores are clustered in and around large cities in each district. As many as 30 stores are located in and around individual larger cities. This makes it easier for the district managers to individually visit each store manager at least once per quarter, as per company policy. Individual store managers may be responsible for all the stores around a city, as many as 30 stores.
- Stores offer month-to-month storage space rental for personal or business use and are a cost-effective and flexible storage alternative. Operating costs are minimized by requiring customers to sign an electronic contract and set up all rentals with electronic payment via either credit or debit card. Payments are made at time of rental contract signing based on a prorated charge for remaining days in the rental month. Full charges apply on the first day of each succeeding month. Customers can cancel their contracts by emptying their storage unit and going online to cancel the electronic contract prior to the first day of the month when the next billing occurs. Per the electronic contract, the last rental month is a full month's charge as there is no prorated charge based on days occupied for the last rental month.
- Tenants rent fully enclosed spaces that can vary in size according to their specific needs and to which they have unlimited, exclusive access. Tenants have responsibility for moving their items into and out of their units. Stores have on-site managers who, with their employees, supervise and run the day-to-day operations, providing tenants with assistance as needed. Self-storage unit sizes vary from 25 square feet to 200 square feet, with an interior heights of 8 feet to 12 feet. The stores are designed with either 100, 200, or 300 units with a standard mix of sizes. The standard unit mix with pricing as of December 31, 2017 is shown in Exhibit 2:

Pistil Storage Case

- Exhibit 2-Standard Unit Size Mix and Pricing Per 100 Storage Units
- Per 100 Units

Standard Quantity	Square Feet		Monthly Price Sq. Foot
10	25	(5 x 5 size)	$1.60
20	50	(5 x 10 size)	$1.50
20	100	(10 x 10 size)	$1.40
20	150	(10 x 15 size)	$1.30
30	200	(10 x 20 size)	$1.20

- Case Assignment

- You are a staff auditor working for the Pistil Storage Inc.'s independent CPA firm, Jacobsen & Jacobsen, which is conducting its first annual audit of the company. Recent prior year audits were done by a "Big 4" audit firm. The date is February 2018 and the firm in working hard on year-end substantive testing.
- One aspect of the audit is to examine the company's analysis of calls received via the toll-free anonymous whistle blower hotline. This is normally done earlier in the audit but this audit work was delayed due to late approval of the Jacobsen & Jacobsen firm to conduct the audit. A review of the hotline activity for the year reveals two hotline calls were received that are concerning. The calls were analyzed by Pistil Storage's internal audit department but no action was taken due to the vague nature of the calls, the fact that no phone numbers were left by the callers, and there is no way to trace the originator of the calls due to the privacy setup on the hot line.

Pistil Storage Case

- The first concerning recorded call occurred in February 2017 and was transcribed as "Hey dudes, I had a hard time finding a number for Pistil Storage but found this number on your website. Why does my 2016 W-2 show 7 months of salary during 2016 when I only worked for 5 months? I am not going to report wages for money I did not receive and I am not going to pay back my unemployment compensation for those 2 months. At least you got the address right. Call me back about this."
- The second concerning recorded call occurred in June 2017 and was "Hello, I just thought you should know that your manager is giving a free month's unit rental in exchange for paying the first month in cash. That is a pain for people like me who don't carry much cash. It seems odd since automatic billing via a credit or debit card is required for the remainder of the contract months. Just saying, because I am mad that I had to go to a bank teller to get the cash so I would qualify for the discount."
- Due to your graduation from a prestigious business school which has a reputation for teaching cutting edge audit theory and practice, the partner in charge of the audit has selected you to apply audit data analytics to see if the company data provides any signals about possible problems in the areas mentioned in the calls. She stated, "This is a shot in the dark assignment but you are one of the best and brightest auditors I know and I am sure you will figure out if we have any problems to deal with."
- Your firm's information technology specialist prepared a data file for you which reflects the 2017 revenue data and various operation metrics for the 1,531 stores. The name of the data file is "Pistil Storage Data 1531 items.csv" A listing of the first 5 records in this data file is shown in Exhibit 3:

- Exhibit 3
- Pistil Storage Revenue Data File With Operation Metrics

GrossRevenue	Location	Units	OccMonths	AdBudget	EqFTEmployees
146100.59	559	100	9.25	5926	5.09
147114.07	168	100	9.31	6892	9.59
511412.79	1453	300	10.79	26829	5.72
361502.19	452	200	11.44	12640	8
336956.99	694	200	10.66	14634	7.33

- GrossRevenue = revenue recorded during 2017 for a specific store location
- Location = specific store number which identifies its location
- Units= number of units at that specific store location
- OccMonths = average number of months unit occupancy during 2017
- AdBudget = 2017 advertising budget for the specific store location
- EqFTEmployees = equivalent full-time employees during 2017 for that store location

- Student Assignment:

- Load the data file and answer the case question via either Excel pivot tables or cluster analysis via R language analysis [or SPSS analysis]. Either approach is acceptable.

Student Case Questions

What is a whistleblower hotline and why do companies have them?

What is the auditor’s responsibility with respect to fraud?

What is the auditor’s responsibility with respect to revenue recognition?

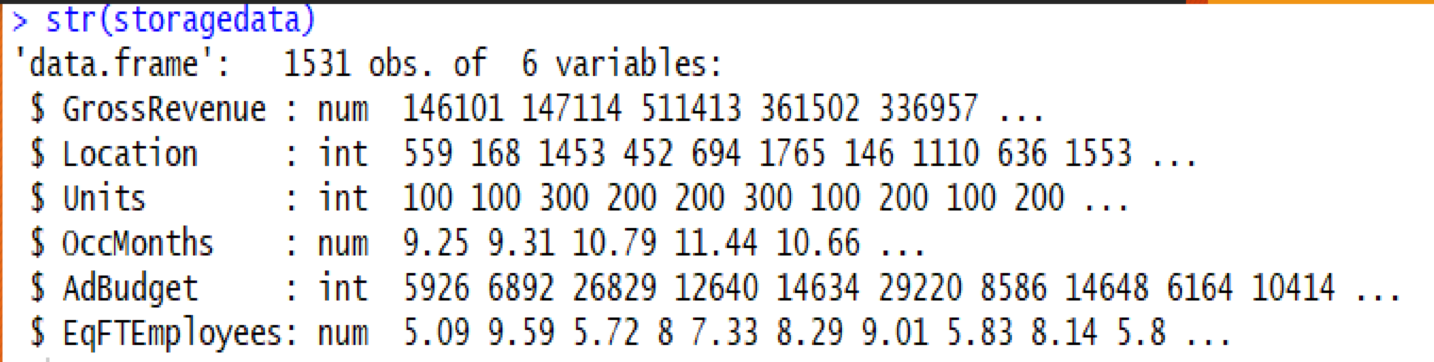
What is the auditor’s responsibility with regard to the two hotline calls?

How does professional skepticism relate to these responsibilities?

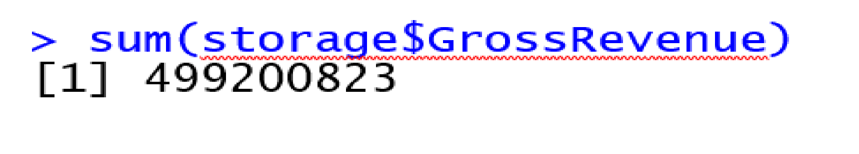
Assume a population of 1,500 revenue locations has 30 locations for which there exist material misstatements. What is the probability of detecting a material misstatement if a sample of one location is randomly selected from the 1,500 locations? Sample selection of 30 locations?

Assume the previous population has been segmented into five equal size subpopulations and one of the subpopulations has been identified as high-risk for the material misstatements. What is the probability of detecting a material misstatement if a sample of 30 locations is randomly selected from the high-risks subpopulation?

Data

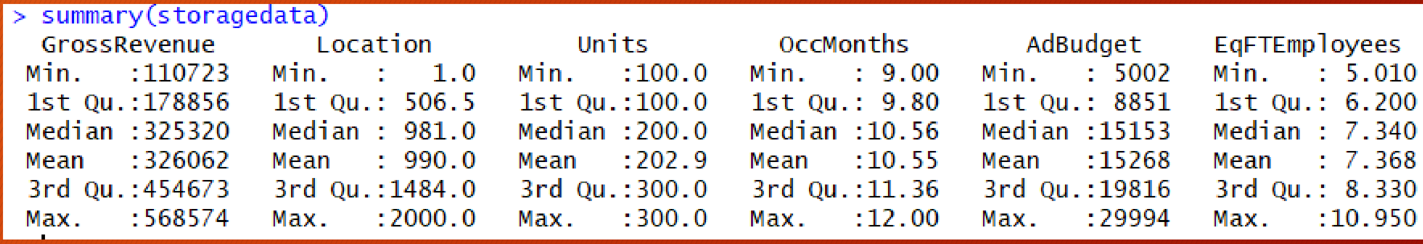


```
> str(storagedata)
'data.frame':   1531 obs. of  6 variables:
 $ GrossRevenue : num  146101 147114 511413 361502 336957 ...
 $ Location     : int  559 168 1453 452 694 1765 146 1110 636 1553 ...
 $ Units        : int  100 100 300 200 200 300 100 200 100 200 ...
 $ OccMonths    : num  9.25 9.31 10.79 11.44 10.66 ...
 $ AdBudget     : int  5926 6892 26829 12640 14634 29220 8586 14648 6164 10414 ...
 $ EqFTEmployees: num  5.09 9.59 5.72 8 7.33 8.29 9.01 5.83 8.14 5.8 ...
```



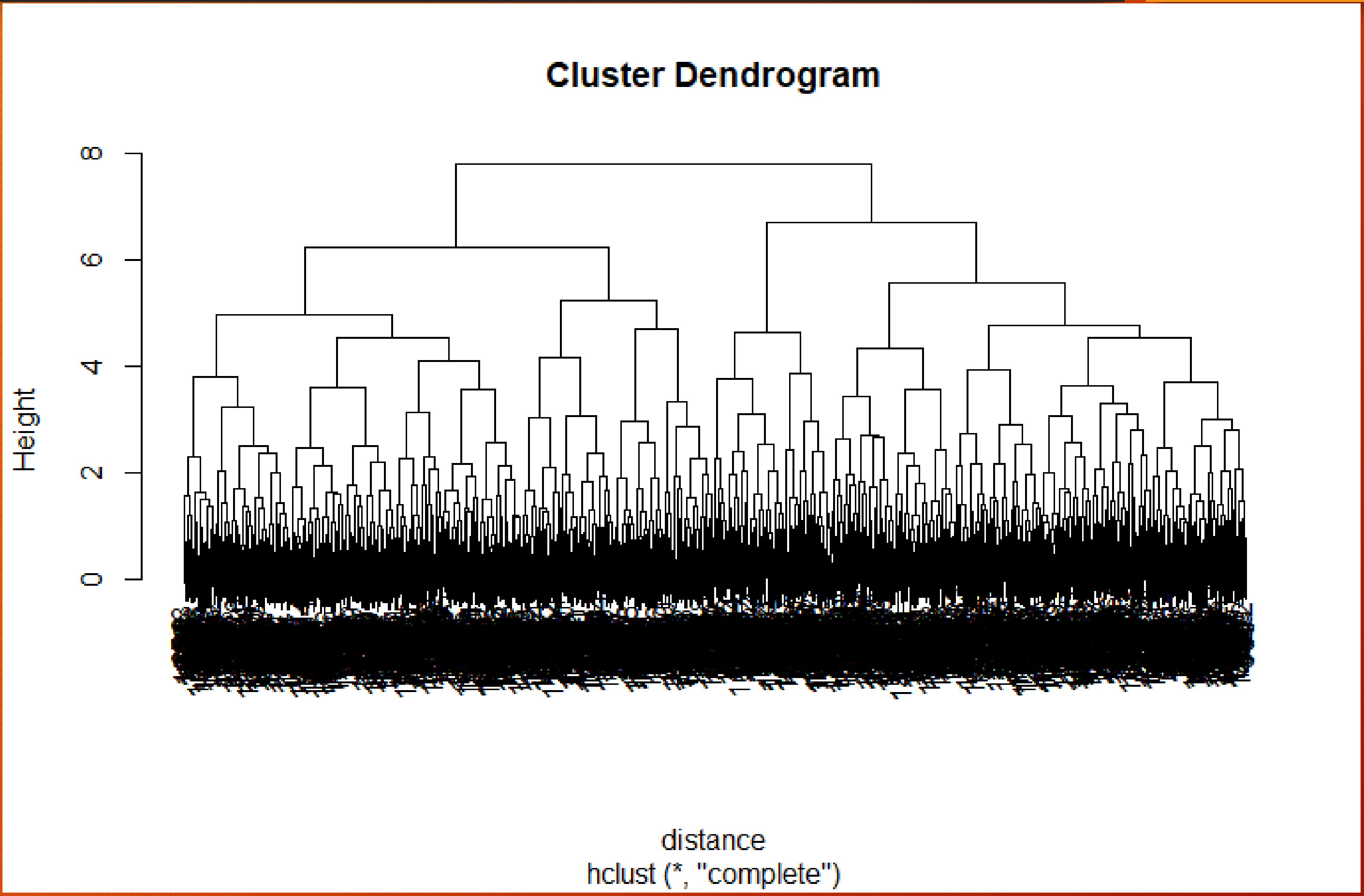
```
> sum(storage$GrossRevenue)
[1] 499200823
```

Sum command is used to add case realism by requiring revenue total be agreed to revenue per general ledger.

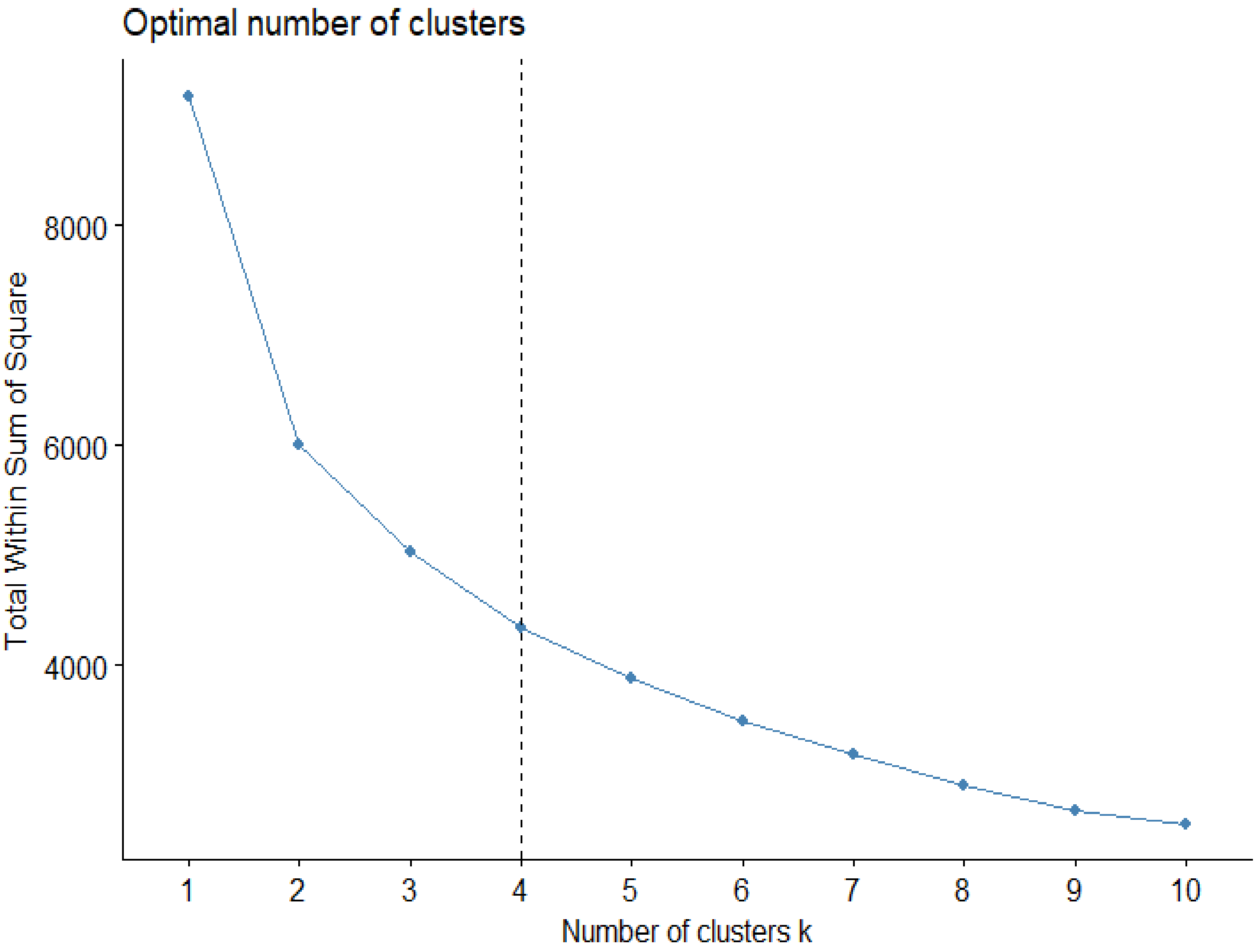


```
> summary(storagedata)
  GrossRevenue       Location          Units          OccMonths        AdBudget      EqFTEmployees   
 Min.   :110723   Min.   :   1.0   Min.   :100.0   Min.   : 9.00   Min.   : 5002   Min.   : 5.010  
 1st Qu.:178856   1st Qu.: 506.5   1st Qu.:100.0   1st Qu.: 9.80   1st Qu.: 8851   1st Qu.: 6.200  
 Median :325320   Median : 981.0   Median :200.0   Median :10.56   Median :15153   Median : 7.340  
 Mean   :326062   Mean   : 990.0   Mean   :202.9   Mean   :10.55   Mean   :15268   Mean   : 7.368  
 3rd Qu.:454673   3rd Qu.:1484.0   3rd Qu.:300.0   3rd Qu.:11.36   3rd Qu.:19816   3rd Qu.: 8.330  
 Max.   :568574   Max.   :2000.0   Max.   :300.0   Max.   :12.00   Max.   :29994   Max.   :10.950  
```

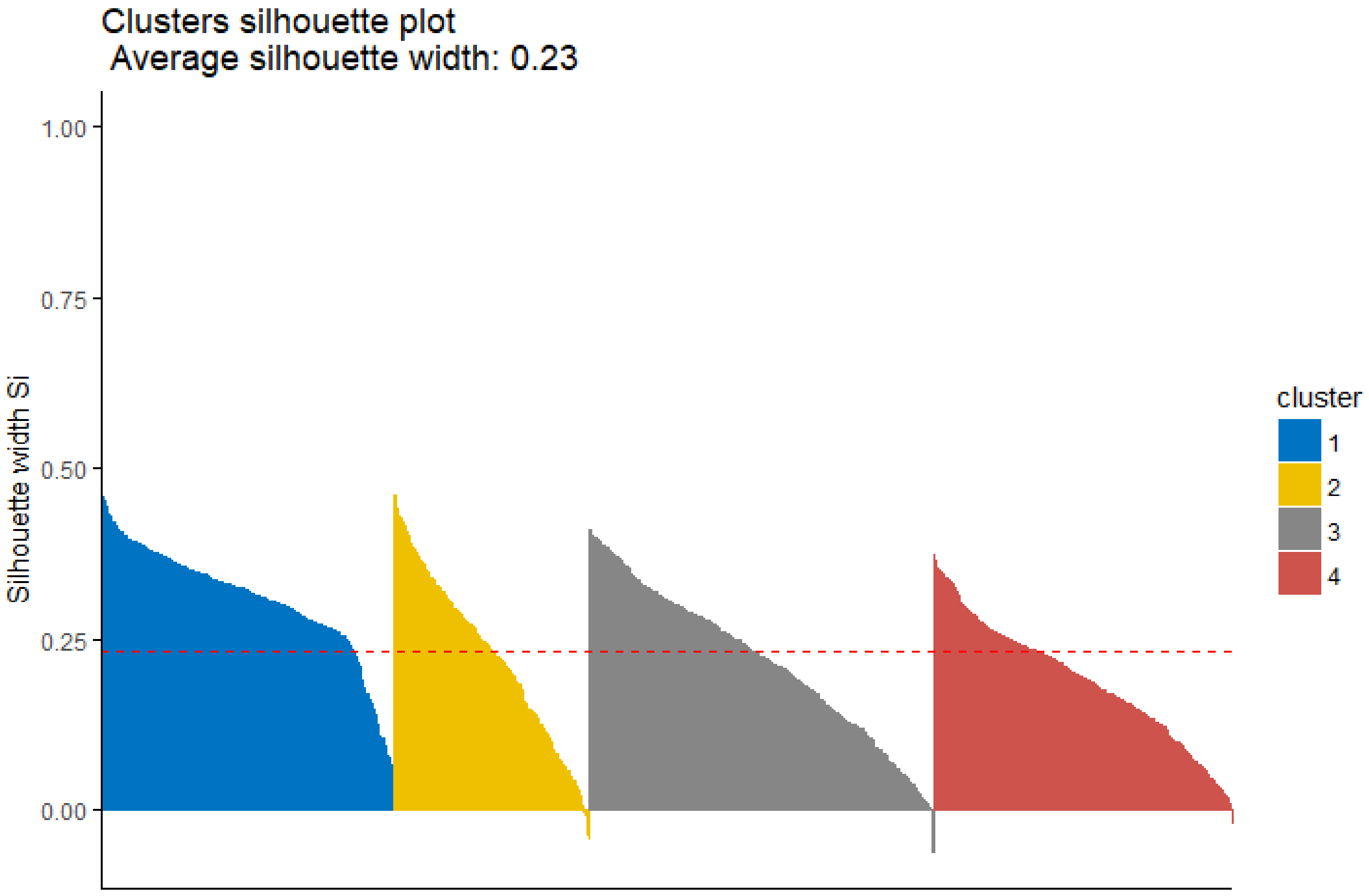
Dendogram—How many clusters?



Elbow Method- How many clusters?

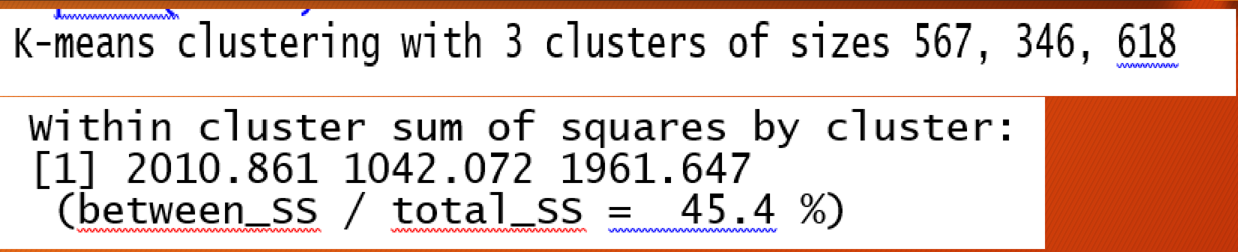


Cluster silhouette Plot

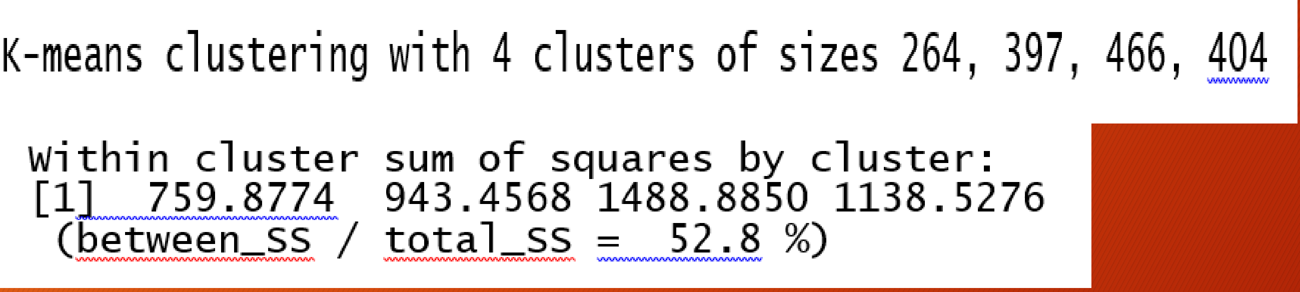


Plot tails falling below 0.00 indicate observations which are misclassified.

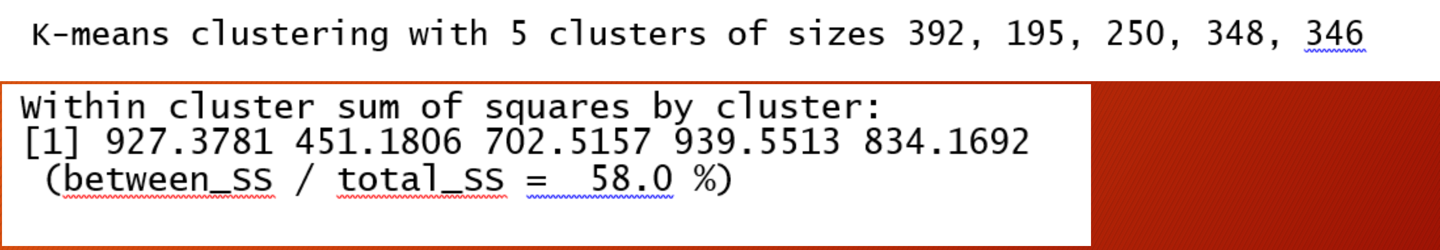
Within Cluster sum of the squares



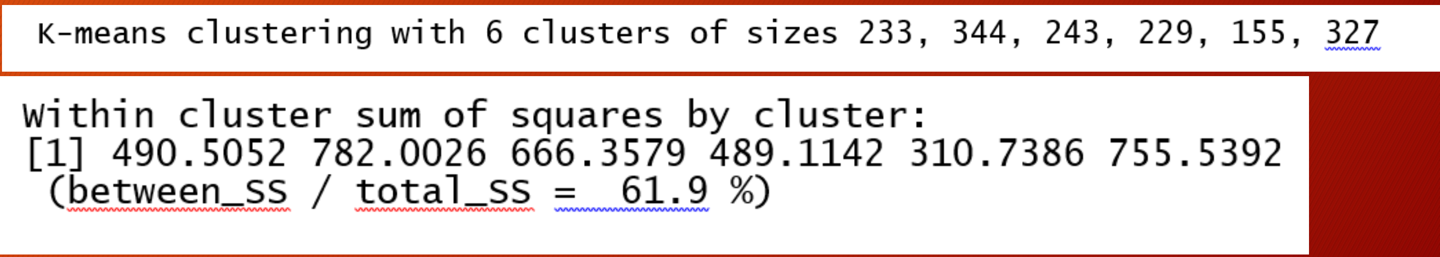
```
K-means clustering with 3 clusters of sizes 567, 346, 618

Within cluster sum of squares by cluster:
[1] 2010.861 1042.072 1961.647
 (between_SS / total_SS =  45.4 %)
```



```
K-means clustering with 4 clusters of sizes 264, 397, 466, 404

Within cluster sum of squares by cluster:
[1]  759.8774  943.4568 1488.8850 1138.5276
 (between_SS / total_SS =  52.8 %)
```

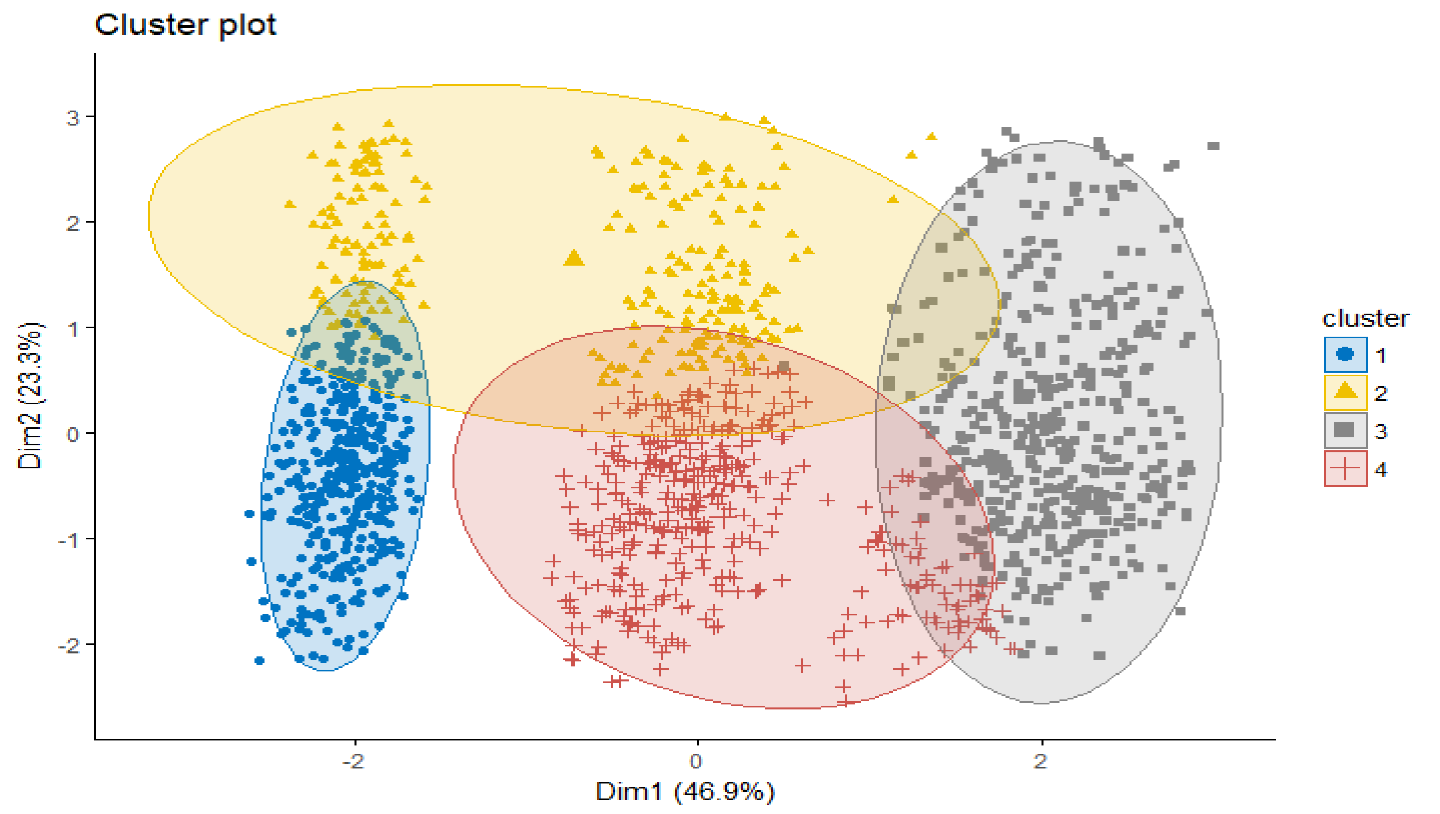


```
K-means clustering with 5 clusters of sizes 392, 195, 250, 348, 346

Within cluster sum of squares by cluster:
[1] 927.3781 451.1806 702.5157 939.5513 834.1692
 (between_SS / total_SS =  58.0 %)
```



```
K-means clustering with 6 clusters of sizes 233, 344, 243, 229, 155, 327

Within cluster sum of squares by cluster:
[1] 490.5052 782.0026 666.3579 489.1142 310.7386 755.5392
 (between_SS / total_SS =  61.9 %)
```

Four cluster plot



This plot indicates some observations don’t group tightly with any of the 4 clusters.

K- Means Values For 4 clusters

- Cluster Sizes

1. 397
2. 264 -- Higher number of employees
3. 466
4. 404 -- Lower occupancy rate

Cluster Means

Cluster	Revenue	LocatioSPSS K-Means Standarized Variablesn	Units	Occupancy	AdBudget	Employees
1	159041	1197	100	10.6	7528	**6.93**
2	270433	297	163	10.6	12032	8.91
3	497021	888	300	10.68	23080	7.44
4	329346	1358	219	**10.41**	15978	6.71

Sort Revenue By Cluster and Examine Revenue By Cluster

Cluster	Gross Revenue	Revenue Per Unit
1	$63,139,277	$1,590
2	$256.952.039	$1,664
3	$86,789,941	$1,658
4	$75,428,030	**$1,507**
AVERAGE		$1,607

Revenue appears significantly lower in cluster 4.

Partial Listing Cluster 1

	GrossRevenue	Location	Units	OccMonths	AdBudget	EqFTEmployees	cluster
1	146100.6	559	100	9.25	5926	5.09	1
9	146835.4	636	100	9.29	6164	8.14	1
12	187690.2	1518	100	11.88	8113	7.92	1
16	123358.0	1715	100	9.81	9738	7.31	1
17	143258.4	1307	100	9.07	7532	7.06	1
19	163333.1	1032	100	10.34	7192	7.78	1
21	151007.5	1587	100	9.56	6495	5.37	1
23	150092.1	623	100	9.50	8125	8.17	1
24	117087.1	1990	100	9.41	7246	6.13	1
36	146588.1	601	100	9.28	8786	6.09	1
38	185654.9	826	100	11.75	9452	5.16	1
45	161285.3	749	100	10.21	5437	6.08	1
47	179969.7	670	100	11.39	8461	8.14	1
51	183828.9	1289	100	11.63	7704	6.57	1
55	148611.1	1805	100	11.41	7987	5.19	1
69	160708.4	1167	100	10.17	9675	5.55	1
73	166754.7	877	100	10.55	8588	5.83	1
79	184792.7	1058	100	11.70	7178	7.87	1
86	151151.4	1261	100	9.57	9763	7.28	1
91	134386.0	1708	100	10.51	6128	5.64	1
99	169082.0	1398	100	10.70	5110	6.31	1
104	178652.8	1187	100	11.31	6447	8.59	1
117	147524.2	1552	100	9.34	6301	5.01	1
128	161837.2	351	100	10.24	7395	7.54	1
131	173010.4	413	100	10.95	5141	5.25	1
133	189358.0	943	100	11.98	7172	8.29	1
145	177699.9	949	100	11.25	5665	5.02	1
149	183283.8	838	100	11.60	5124	8.04	1
150	14						

R Language K-Means Findings

- Cluster 4 has slightly less occupancy per unit and significantly less revenue per unit
  - Possible diversion of revenue via cash payments?
  - [Note: This is same finding for SPSS Cluster 1 at end of this presentation]

- Cluster 2 has significantly more full-time equivalent employees
  - Possible diversion of payroll funds?
  - [Note: This is same finding for SPSS Cluster 4 at end of this presentation]

  - What do these clusters correspond to ?

Clusters Categorized by Store District

Custer	Southeast 1-400	Midwest 401-800	Northeast 801-1200	Southwest 1201-1600	West 1601-2000
1	7	89	106	97	98
2	200	50	14	0	0
3	99	111	113	99	44
4	0	61	85	103	155

Pistil Case Audit Recommendations

[Alternatively: testing can be initiated by internal audit staff and reviewed by external auditor]

1. Test sample of occupancy contracts for West Region
   - Determine recorded months occupancy
   - Compare recorded months occupancy to revenue recorded for contract
   - Call or contact customers to see if cash payments made that have not been recorded as revenue.
   - If fraud indicated, discuss with partner about having internal audit test 100% of West Region contracts for incorrect occupancy months

2. Test sample of payroll records of terminated employees for Southeast Region
   - Validate number of months worked per payroll records by contacting employees and/or state unemployment benefits agency
   - If discrepancies exist, examine payroll check deposits for months after termination to determine who deposited them
   - If fraud indicated, discuss with partner about having internal audit test 100% of Southeast Region payroll terminations for incorrect month of termination and diversion of funds.

Case Learning Objectives

- Develop student critical thinking skills in an audit setting
- Increase student understanding of auditor revenue recognition responsibilities
- Increase student understanding of auditor due diligence responsibilities
- Increase student understanding of auditor fraud responsibilities
- Increase student understanding of possible audit value from unsupervised learning
- Increase student facility with R, SPSS, or Excel Pivot tables

Future Case Revisions???

- Add random revenue sampling assignment as precursor to cluster analysis assignment.

- Include auditor internal control chart analysis from prior audit

- Add additional variables to data files, for example, individual manager employee IDs for each location.

- Add requirement to create regression model to forecast revenue by locations and compare to actual forecast. Compare the results of this analysis to results of cluster analysis.

References

- Kassambara, A. 2017. Practical Guide To Cluster Analysis in R : Unsupervised Machine Learning. Published by STHDA.

- McKee, T.E. 2021. “Analyzing and Audit Population via Either Excel Pivot Tables and/or R Language Cluster Analysis.” ***Current Issues in Auditing***. Vol 15, No. 1 , Spring issue.

- McKee, T.E. 2021 “Generating Synthetic Data For Improved Accounting/Auditing Courses.” Unpublished paper under development.

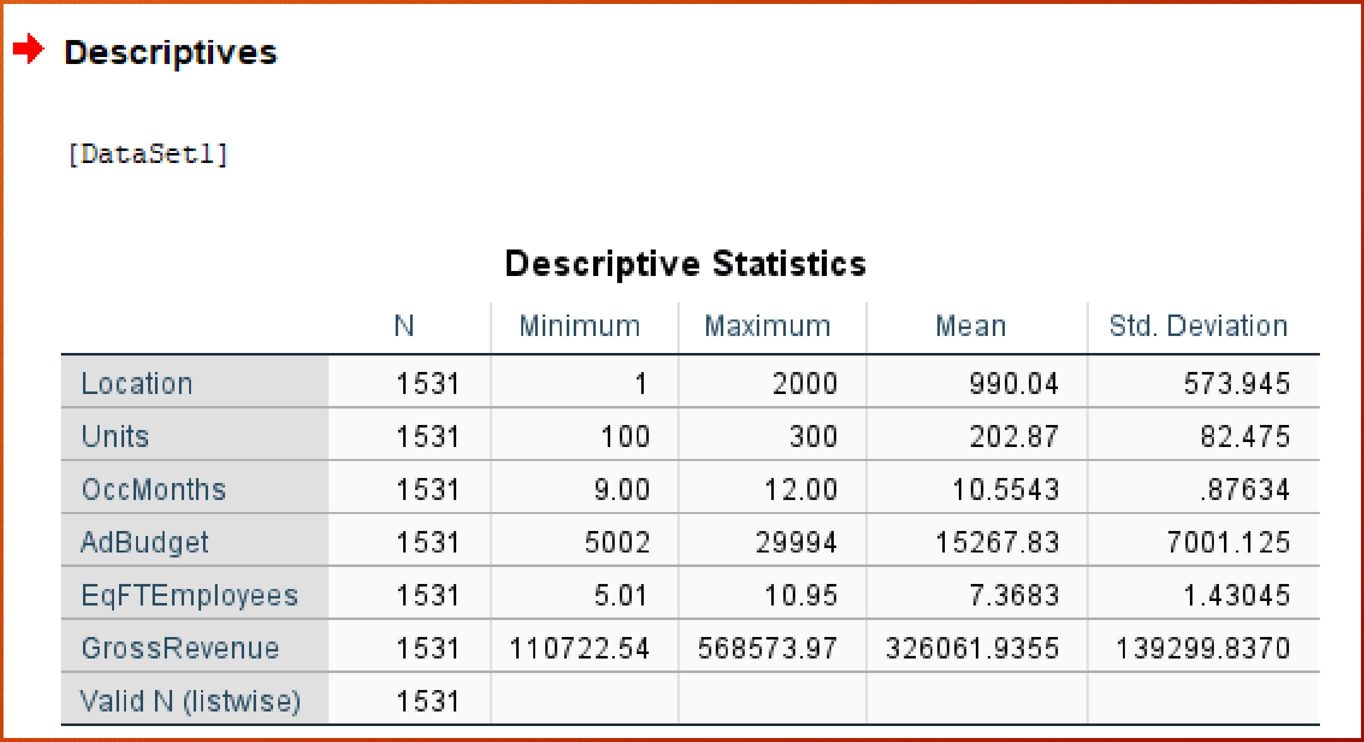
- Templ, M. 2016. *Simulation For Data Science With R*. Packt Publishing Ltd., Birmingham, U.K

- Thiprungsri, S. and M.A. Vasarhelyi. 2011. Cluster Analysis for Anomaly Detection. *The International Journal of Digital Accounting Research*. Vol. 11, pp. 69-84.

- Van der Schaar, M. and N. Maxfield, September 7, 2020. Synthetic Data: Breaking The Data Logjam In Machine Learning For Healthcare. https://www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/

SPSS Descriptives For Data Standardization

Standardize variables: Analyze→Descriptive Statistics→Descriptives→ Save Standardized Values



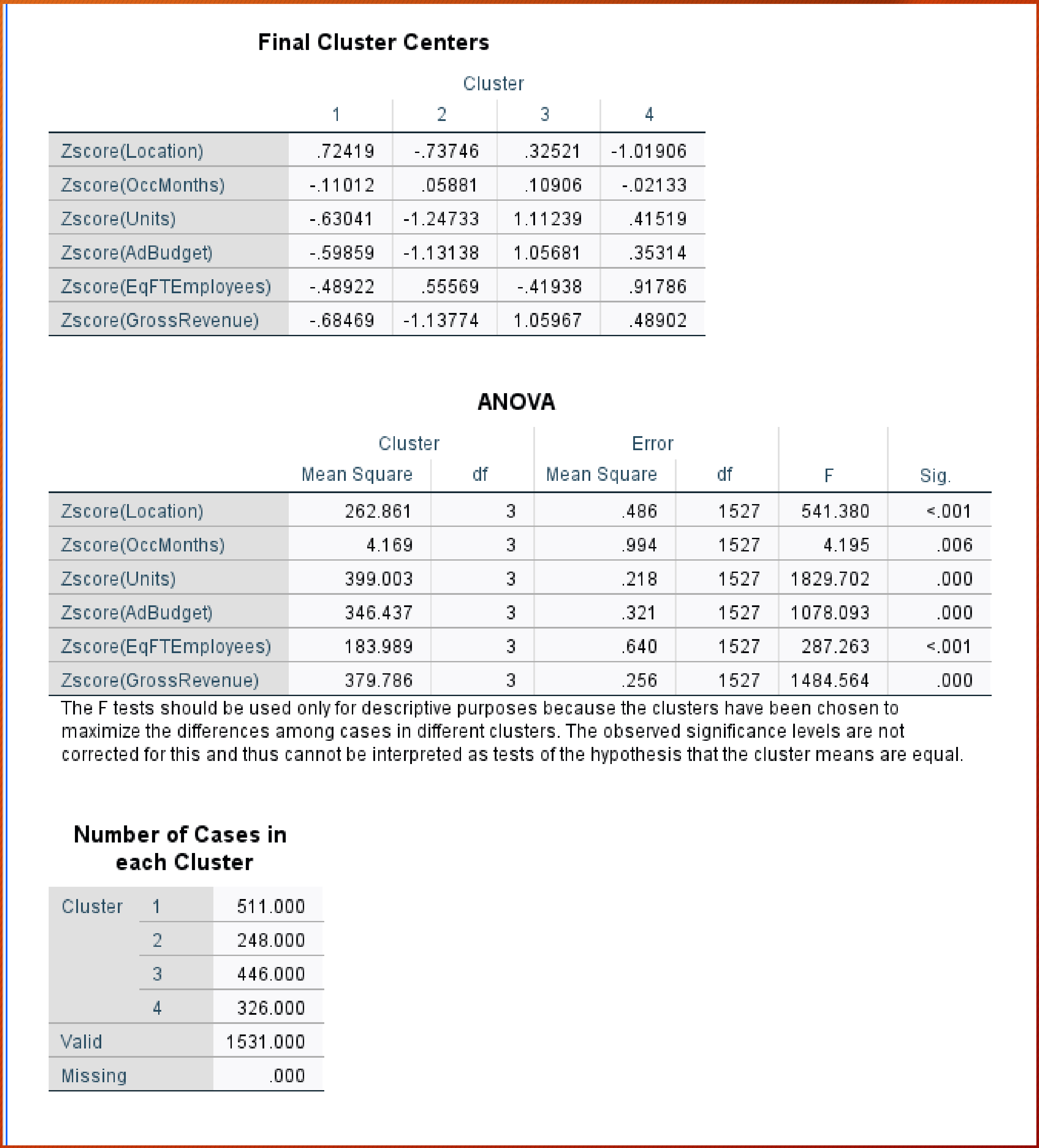
Descriptives

```
[DataSet1]
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Location	1531	1	2000	990.04	573.945
Units	1531	100	300	202.87	82.475
OccMonths	1531	9.00	12.00	10.5543	.87634
AdBudget	1531	5002	29994	15267.83	7001.125
EqFTEmployees	1531	5.01	10.95	7.3683	1.43045
GrossRevenue	1531	110722.54	568573.97	326061.9355	139299.8370
Valid N (listwise)	1531				

SPSS Cluster Analysis Standardized Variables



Final Cluster Centers

	Cluster			
	1	2	3	4
Zscore(Location)	.72419	-.73746	.32521	-1.01906
Zscore(OccMonths)	-.11012	.05881	.10906	-.02133
Zscore(Units)	-.63041	-1.24733	1.11239	.41519
Zscore(AdBudget)	-.59859	-1.13138	1.05681	.35314
Zscore(EqFTEmployees)	-.48922	.55569	-.41938	.91786
Zscore(GrossRevenue)	-.68469	-1.13774	1.05967	.48902

ANOVA

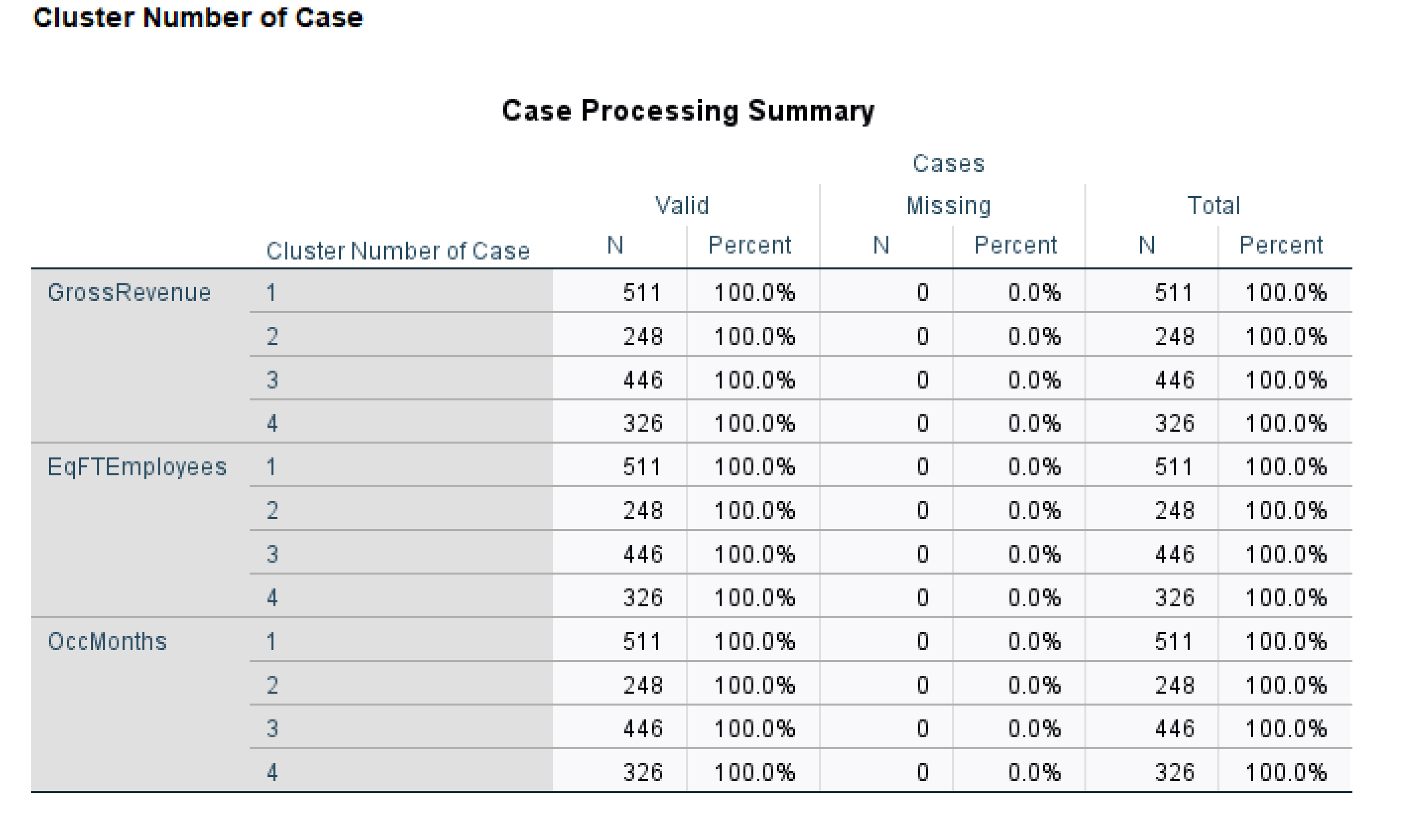
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore(Location)	262.861	3	.486	1527	541.380	<.001
Zscore(OccMonths)	4.169	3	.994	1527	4.195	.006
Zscore(Units)	399.003	3	.218	1527	1829.702	.000
Zscore(AdBudget)	346.437	3	.321	1527	1078.093	.000
Zscore(EqFTEmployees)	183.989	3	.640	1527	287.263	<.001
Zscore(GrossRevenue)	379.786	3	.256	1527	1484.564	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	511.000
	2	248.000
	3	446.000
	4	326.000
Valid		1531.000
Missing		.000

SPSS K-Means Standardized Variables



Cluster Number of Case

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Cluster Number of Case	N	Percent	N	Percent	N	Percent
GrossRevenue	1	511	100.0%	0	0.0%	511	100.0%
	2	248	100.0%	0	0.0%	248	100.0%
	3	446	100.0%	0	0.0%	446	100.0%
	4	326	100.0%	0	0.0%	326	100.0%
EqFTEmployees	1	511	100.0%	0	0.0%	511	100.0%
	2	248	100.0%	0	0.0%	248	100.0%
	3	446	100.0%	0	0.0%	446	100.0%
	4	326	100.0%	0	0.0%	326	100.0%
OccMonths	1	511	100.0%	0	0.0%	511	100.0%
	2	248	100.0%	0	0.0%	248	100.0%
	3	446	100.0%	0	0.0%	446	100.0%
	4	326	100.0%	0	0.0%	326	100.0%

SPSS K-Means Standardized Variables Means

Cluster/ # Cases	Gross Revenue/ Revenue Per Unit	Location	Units	Occupancy	AdBudget	Employees
1 511	230685 **1528**	1405	151	10.46	11077	**6.67**
2 248	167575 1676	567	100	10.61	7347	8.16
3 446	473674 1617	1177	293	10.65	22667	6.77
4 326	394183 1699	405	232	**10.54**	17179	**8.68**

Note: Since K-Means Cluster Analysis involves random starting points for forming the initial clusters, you won’t get the same clusters unless you use the same random starting point which R Language analysis and SPSS analysis did NOT.

However, SPSS also identified a cluster with lower occupancy and lower revenue per location as well as a cluster with a higher number of employees per location.