The Art of Employing Data Science to Improve Cost Data Analysis

By Greg Wiegand, Shavaiz Saood, and Richard Shea

International Cost Estimating and Analysis Association
2018 Professional Development and Training Workshop
Phoenix, Arizona

June 2018

Table of Contents

Table of Tables

Table of Figures

Author Biographies

Greg Wiegand has over five years of experience supporting Navy, Marine Corps and Department of Energy (DOE) clients. Mr. Wiegand holds a Bachelor of Science in Operations and Supply Chain Management and Decision Sciences from the University of Dayton. He currently supports the Naval Sea Systems Command (NAVSEA) Cost Engineering and Industrial Analysis (05C) Cost Analysis Team, specifically estimating costs for the DDG 51, a class of missile destroyer.

Shavaiz Saood joined Herren Associates in 2015 with a bachelor in Finance from Duquesne University. Shavaiz currently supports NAVSEA 05C in the areas of cost engineering and data analysis. He specializes in the development of Life Cycle Cost Estimates (LCCEs), Independent Cost Estimates (ICE), Business Case Analyses (BCAs), along with other estimates and analyses. Prior to Herren, Shavaiz worked in the finance sector as an Investment Operations Analyst at the Bank of New York Mellon.

Richard Shea joined Herren Associates in 2016 as a lead Senior Cost Analyst supporting NAVSEA 05C. Mr. Shea has years of experience supporting various Department of Defense (DoD) Organizations from the DoD Health Affairs, Defense Health Agency (DHA), Department of Veterans Affairs, US Marine Corps, US Army, and now the US Navy. Mr. Shea is experienced in building life cycle cost estimates, business case analysis, and independent cost estimates supporting DoD systems and services.

Company

Herren

Founded in 1989, Herren Associates is an engineering and management consulting firm with a proven record of maximizing the value of every taxpayer dollar. As trusted advisors to federal executives, we partner with clients to drive operational improvements and manage performance - maximizing efficiency and cost effectiveness.

Paper Abstract

Preface

While large data sets are highly desired and used frequently, there is not a universally accepted standard format for large data extractions and transfers. This presentation focuses on gaining insights into large amounts of data, spotting inconsistencies, and transforming data into a usable format. This presentation examines the possibility of putting data into a standard database format so it can be easily manipulated and cross compared against other datasets that may have comparable elements.

In the field of cost estimating, large data sets are highly desired and used frequently. Some would say data is the lifeblood of a good cost estimate. There is not a universally accepted standard format for large data extractions and transfers, so obtaining data from ten different sources usually means obtaining data in ten different formats. Within specific communities, data formats may be similar, but what happens when the data you've long awaited finally arrives in a format you don't understand, is seemingly unusable, or simply illogically formatted? This study focuses on using data science to gain better insights into large amounts of data, spotting inconsistencies, as well as getting data into a usable format. In addition, this study examines the possibility of putting data into a standard database format so it can be easily manipulated and cross compared against other datasets that may have comparable elements. In this specific case, a large Enterprise Resource Planning (ERP) data extraction was transferred into a database format, then cross compared with Common Systems Engineering (CSE) 7300 and Task Planning Sheet (TPS) data, as well as claims made in a Government Furnished Equipment (GFE) model. The transformation of this data allowed analysts to garner valuable insights into subtle inconsistencies and improve the accuracy of the estimate.

Data Science Discussion

"Data are the foundation of every cost estimate." (GAO, 2009)

Characteristics of a good cost estimate are accuracy, credibility, and defensibility of the delivered product. The quality of data supporting a cost estimate is reflected in the estimate's credibility. One would ask; 'Does this make sense?'. High quality data will lead to more informed decisions where poor data will either be disregarded or create poor decisions. Historical data is the backbone of a good estimate and good data provides credibility, accuracy, and defensibility.

The availability of valid data is one of the basic characteristics of a credible cost estimate. An estimator must interview data sources and document any relevant information to identify the data veracity. Documenting data sources will allow the estimate to be auditable and traceable to the cost element. This is especially necessary if there is a lack of quality data available.

The cost analyst will need to document any normalization methodologies applied to the data. Cost Estimators must be able to discern data quality by investigating the reliability and accuracy of the data.

Data collection is a top priority for cost estimators and data can come from many sources. Data sources can be internal to the organization such as historical costs on similar systems or external to the organization such as costs for raw material, supplies, and components.

Assessment of the data quality is critical for determining the applicability and usability of the data elements. The context with which we source data and the contextual completeness is of the utmost importance. The presence of data anomalies is important because it reduces the database utility and thereby its effectiveness. Good contextual completeness will have specifics of technical and programmatic attributes identified to the data.

Data quality is more than just data accuracy. Data quality is addressed in different areas like completeness, consistency, and currency. The data must tell a consistent comprehensive story each time. The data being used in the analysis must be applicable to the question at hand.

Data can be either quantitative or qualitative as illustrated in *Table 1: Quantitative or Qualitative Data Types* below. The analysis will require different approaches to manipulate and use the appropriate data. Quantitative data is much more easily used and manipulated where qualitative data will need to have metrics around the quality to place a comparative measure around data elements that can be analyzed.

Quantitative	Qualitative
Mass	Manufacturability
Length	Complexity
Velocity	Quality
Weight	Aesthetics

Table 1: Quantitative or Qualitative Data Types

Also important to cost estimating is the nature of data (types, formats, stories). An effective data story is much more than a scatter plot from Excel®, but a well thought out insight of what the data is telling the analyst. Data needs to be in a format that can be stored, retrieved, and analyzed. Structured data types are more easily manipulated than unstructured data types, but both can be useful in the hands of a skilled analyst.

True data completeness means the data has the necessary data elements to lead the analyst to the appropriate conclusion. Missing or incomplete data can be a challenge to the analyst since it takes special knowledge to fill in the missing components or elements. Knowing the data source will go a long way to fill in data gaps. Relying on a knowledgeable subject matter expert (SME) can assist the analyst to fill in the missing pieces.

Data can be structured, semi-structured, and unstructured. Unstructured data is not always conducive to analysis as it is hard to manipulate, work with, and glean important information.

Table 2: Primary or Secondary Data Types below illustrates primary and secondary data sources:

Basic Primary and Secondary Data Sources (GAO Cost Estimating Guide)

Data type	Primary	Secondary
Basic accounting records	x	
Data collection input forms	x	
Cost reports	x	x
Historical databases	x	x
Interviews	x	x
Program briefs	x	x
Subject matter experts	x	x
Technical databases	x	x
Other organizations	x	x
Contracts or contractor estimates		x
Cost proposals		x
Cost studies		x
+Focus groups		x
Research papers		x
Surveys		x

Table 2: Primary or Secondary Data Types

A high volume of data can be as much of a challenge as not enough data. The analyst needs to know how to filter and manipulate the data to uncover the underlying story and convey that to the audience. Too much data can be just as hard to use as too little data, the analyst needs to know how to filter and manipulate to find what they're seeking. (ICEAA, 2013)

Problem Statement

We acquired a large data extract from an Enterprise Resource Planning (ERP) database through a United State Government Program Office (PO) representative. This data extract was supporting

an analysis effort requested by the PO. The data we received was geared towards tracking funding for various activities for execution by project through the Planning, Programming, Budgeting and Execution (PPBE) process.

Specifically, the ERP file contained data on the project planned cost, budget, commitments, obligations, actual costs, actual revenues, assigned costs, and available budget by fiscal year (FY). This database contained over 6000 rows of data.

There were unique problems we had to solve for this project:

1. The ERP database was organized in a contract execution hierarchy that contained one or multiple ships at the first level indenture in a parent child relationship. With an inconsistent naming convention.
2. Single or multiple children would sum to the parent, with lower level parent(s) summing to a higher-level parent. Useful data nestled in with extraneous data.
3. The Work Breakdown Structure (WBS) was ten-levels deep and was in an unusable format when we first received it and was something they took the time to collect but had no way of extracting any usable insights.
4. ERP Data was for tracking funding execution through the Planning, Programming, Budgeting, and Execution (PPBE) process
5. We also support the Government Furnished Equipment (GFE) estimate, which goes through annual decision reviews. Prior to this data, the Navy customer didn't have a robust approach of assessing their estimate against historical actuals.

The first task we needed to accomplish was to scrub the data to identify the child elements that rolled up into parent elements using Excel® functions

1. The function required us to look at the WBS numbering to identify the number of children for each element
2. Children were identified as those without any sub elements
3. This helped in allowing for a meaningful pivot to be constructed

Here in *Table 3: Source Naming Convention* is an example of the naming convention and data organizational structure that we found upon data delivery. We can see that the numbering system contained difficulty in developing a consistent and reliable table format system. We could identify the Parent/Child relationship where children were basic data elements and the parents were summations of children and other subordinate parents. Where in *Table 4: Source WBS Naming Description* provides the level where data elements are described.

ERP WBS	Count of Children	Parent/Child
11	1443	Parent
11	1443	Parent
1101	1325	Parent
110101	16	Parent
11010101	9	Parent
1101010101	1	Parent
110101010101	0	Child
1101010102	1	Parent
110101010201	0	Child
1101010103	4	Parent
110101010301	0	Child
110101010302	0	Child
110101010303	0	Child
110101010304	0	Child
11010102	5	Parent
1101010201	2	Parent
110101020101	0	Child
110101020102	0	Child
1101010202	1	Parent
110101020201	0	Child

Table 3: Source Naming Convention

WBS Desc Lvl 3	WBS Desc Lvl 4	WBS Desc Lvl 5	WBS Desc Lvl 6	WBS Desc Lvl 7
Program Element	Hardware or System Element	Ship Description	Performer	Task or Work to be Performed

Table 4: Source WBS Naming Description

Important Things to consider

Planned Costs and Budget were not consistently applied at the child level elements and these dollars do not equal the Assigned Costs in the raw data. We decided that these costs were for planning and budgeting purposes and did not consistently reflect the dollar amounts found in the Assigned Costs in the ERP database.

We found that the Commitments plus Obligations plus Actual Costs did equal the assigned costs. Since these data elements map out to the child level elements, we scrubbed the data to make sure they added up to what's in the parent or roll up elements. We were confident that the Assigned Costs were the maximum actual costs that would or will be expended for the particular task assigned. We considered these to be the costs at the executed CLIN level.

Commitments + Obligations + Actual Costs = Assigned Costs

Right now, the pivot is filtered to show only the Electronics and Ordnance pieces of the database. During the analysis, we have gone through the database to identify ships and ships systems at the child element level. At the conclusion of the analysis, we accomplished a complete mapping of ships and systems to begin better understanding the data and cross checking per ship costs and per system costs.

Our first data table sample, *Table 5: First Raw Data Sample*, illustrates the basic parent child relationship arrangement we were finding. This sample is for a government supplied contract field service that was done by a single Government Performer and applied to support two ships. The Direct Cite and Reimbursable indicate that the Government Performer did some of the work and contracted other work to one or more sub-contractors. There was not enough information in the database to identify the sub-contractors but the Government Performer was identified. The *Table 6: Expanded First Raw Data Sample*, illustrates how we extracted the data into a more useable format in our analysis.

As you can see in *Table 7: Second Raw Data Sample*, a second example of challenges we met where there were parents and child elements that were added together into a higher parent level. This provided challenges where it was difficult to readily identify the data elements that make up a higher parent element. We could accomplish this as illustrated in Table 8: Expanded Second Raw Data Sample, where the higher-level parent contains the costs of lower level parents and children. The lower level parents having their own children.

First Raw Data Table Sample

Parent/Child	Description	Currency Key	Planned Cost	Budget	Commitments	Obligations	Actual Costs	Actual Revenues	Assigned Costs	Available Budget	Year
Parent	Fiscal Year CONTRACT FIELD SERVICES	USD	-	-	200,000	104,229	1,374,696	-	1,678,925	(1,678,925)	FY11
Parent	System #1 Name - CONTRACT FIELD SVCS Ship #1	USD	-	-	200,000	40,196	1,013,897	-	1,254,093	(1,254,093)	FY11
Parent	Government Performer #1	USD	-	-	200,000	40,196	1,013,897	-	1,254,093	(1,254,093)	FY11
Child	DIRECT CITE	USD	-	-	200,000	34,636	999,499	-	1,234,135	(1,234,135)	FY11
Child	REIMBURSABLE	USD	-	-	-	5,560	14,398	-	19,958	(19,958)	FY11
Parent	CONTRACT FIELD SVCS Ship #2	USD	-	-	-	64,033	360,799	-	424,832	(424,832)	FY11
Parent	Government Performer #1	USD	-	-	-	64,033	360,799	-	424,832	(424,832)	FY11
Child	DIRECT CITE	USD	-	-	-	64,033	335,074	-	399,107	(399,107)	FY11
Child	REIMBURSABLE	USD	-	-	-	-	25,725	-	25,725	(25,725)	FY11

Table 5: First Raw Data Sample

First Raw Data Table Expanded Sample

	Data Provided			Data Extracted				
Parent/Child	Description	Assigned Costs	Year	Ship	Electronics and Ordnance	Performer	DIRECT CITE	REIMBURSABLE
Parent	Fiscal Year CONTRACT FIELD SERVICES	1,678,925	FY11	#1	System 1	Contract Field Svcs		
Parent	System #1 Name - CONTRACT FIELD SVCS Ship #1	1,254,093	FY11	#1	System 1	Contract Field Svcs		
Parent	Government Performer #1	1,254,093	FY11	#1	System 1	Government Performer #1		
Child	DIRECT CITE	1,234,135	FY11	#1	System 1	Government Performer #1	DIRECT CITE	
Child	REIMBURSABLE	19,958	FY11	#1	System 1	Government Performer #1		REIMBURSABLE
Parent	CONTRACT FIELD SVCS Ship #2	424,832	FY11	#2	System 1	Government Performer #1		
Parent	Government Performer #1	424,832	FY11	#2	System 1	Government Performer #1		
Child	DIRECT CITE	399,107	FY11	#2	System 1	Government Performer #1	DIRECT CITE	
Child	REIMBURSABLE	25,725	FY11	#2	System 1	Government Performer #1		REIMBURSABLE

Table 6: Expanded First Raw Data Sample

Second Raw Data Table Sample

Parent/Child	Description	Currency Key	Planned Cost	Budget	Commitments	Obligations	Actual Costs	Actual Revenues	Assigned Costs	Available Budget	Year
Parent	System HARDWARE Ship #3	USD	-	-	17,000	73,115,793	60,132,597	-	133,265,390	(133,265,390)	FY13
Child	Government CONTRACTS	USD	-	-	-	72,638,078	59,679,883	-	132,317,960	(132,317,960)	FY13
Parent	Government Performer #2	USD	-	-	-	437,072	424,617	-	861,689	(861,689)	FY13
Child	REIMBURSABLE	USD	-	-	-	26,381	225,871	-	252,252	(252,252)	FY13
Child	DIRECT CITE	USD	-	-	-	410,691	198,746	-	609,437	(609,437)	FY13
Parent	Government Performer #3	USD	-	-	-	30,127	28,097	-	58,224	(58,224)	FY13
Child	REIMBURSABLE	USD	-	-	-	30,127	28,097	-	58,224	(58,224)	FY13
Parent	Government Performer #4	USD	-	-	17,000	10,517	-	-	27,517	(27,517)	FY13
Child	REIMBURSABLE	USD	-	-	-	10,517	-	-	10,517	(10,517)	FY13
Child	DIRECT CITE	USD			17,000	-	-	-	17,000	(17,000)	FY13

Table 7: Second Raw Data Sample

Second Data Table Results

		Data Provided			Data Extracted				
Parent/Child	Description	Assigned Costs	Year	Ship	Electronics and Ordnance	Performer	DIRECT CITE	REIMBURSABLE
Parent	System HARDWARE Ship #3	133,265,390	FY13	Ship #3	System 2	Hardware		
Child	Government CONTRACTS	132,317,960	FY13	Ship #3	System 2	Gov't Contracts		
Parent	Government Performer #2	861,689	FY13	Ship #3	System 2	Government Performer #2		
Child	REIMBURSABLE	252,252	FY13	Ship #3	System 2	Government Performer #2		REIMBURSABLE
Child	DIRECT CITE	609,437	FY13	Ship #3	System 2	Government Performer #2	DIRECT CITE	
Parent	Government Performer #3	58,224	FY13	Ship #3	System 2	Government Performer #3		
Child	REIMBURSABLE	58,224	FY13	Ship #3	System 2	Government Performer #3		REIMBURSABLE
Parent	Government Performer #4	27,517	FY13	Ship #3	System 2	Government Performer #4		
Child	REIMBURSABLE	10,517	FY13	Ship #3	System 2	Government Performer #4		REIMBURSABLE
Child	DIRECT CITE	17,000	FY13	Ship #3	System 2	Government Performer #4	DIRECT CITE	

Table 8: Expanded Second Raw Data Sample

Problem Solution

Parent/Child	(All)
Row Labels	**Sum of Assigned Costs**
FY13	**$267,478,210**
Ship #3	**$267,478,210**
System 2	$267,478,210
Government Performer #1	**$132,317,960**
Government CONTRACTS	$132,317,960
Government Performer #2	**$1,723,378**
DIRECT CITE	$609,437
Government Performer #2	$861,689
REIMBURSABLE	$252,252
Government Performer #3	**$116,448**
Government Performer #3	$58,224
REIMBURSABLE	$58,224
Government Performer #4	**$55,034**
DIRECT CITE	$17,000
Government Performer #4	$27,517
REIMBURSABLE	$10,517
Hardware	**$133,265,390**
System HARDWARE Ship #3	$133,265,390
Grand Total	**$267,478,210**

Table 9: Pivot Table Sample from First Data Sample

Parent/Child	Child
Row Labels	**Sum of Assigned Costs**
FY13	**$133,265,390**
Ship #3	**$133,265,390**
System 2	$133,265,390
Government Performer #1	**$132,317,960**
Government CONTRACTS	$132,317,960
Government Performer #2	**$861,689**
DIRECT CITE	$609,437
REIMBURSABLE	$252,252
Government Performer #3	**$58,224**
REIMBURSABLE	$58,224
Government Performer #4	**$27,517**
DIRECT CITE	$17,000
REIMBURSABLE	$10,517
Grand Total	**$133,265,390**

Table 10: Pivot Table from Second Data Sample

Problem Solution Approach

The ERP database was organized in a contract execution hierarchy that may contain one or multiple ships at a higher-level indenture in a parent to child relationship. Lower level parent(s) would sum to a higher-level parent. Single or multiple children could sum to a lower level parent. Commonly, there would be a lower level parent and child summing to a higher-level parent. The child element was the base dollar where the parent element always has a sub element where it would be another parent or a child. The structure of the data hierarchy made it difficult to identify the children to the higher-level parents.

The initial step to decode the ERP database was to identify costs at the child level that are associated with a specific ship hull number. Many times, parent costs would contain costs for multiple ships and then be broken down into child element(s) to a unique single ship hull number.

The next step of the ERP database analysis tasking required the team to shred out multiple bits of information that were contained in a higher-level parent data field and sharing that down to the child elements. The higher-level parent data field may contain one or more of the following: electronics system, ordnance system, ship number, tasking, and/or performer. The lower level parent and children fields would contain additional information down to direct cite and

reimbursable. We utilized the parent child relationship discussed above to assign the data description associated with the data element and cost down to the child level.

Examples of the Results

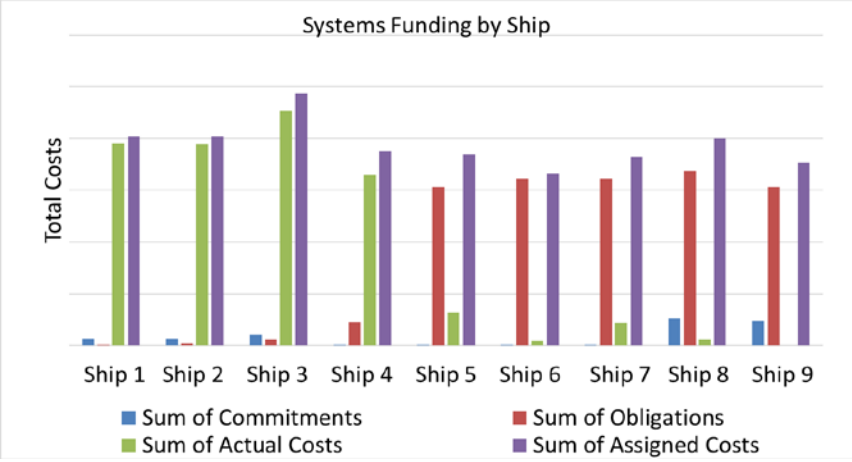

Figure 1: System Funding by Ship

As you can see with in *Figure 1: System Funding by Ship*, the total costs of the ships are comparable and one can see that first four ships are complete or nearing completion where the later five ships are just getting planned and started. Further study should be place into Ship 3 of why there is an increase in this ship. It could be a new weapon or ordnance system that is being installed and tested.

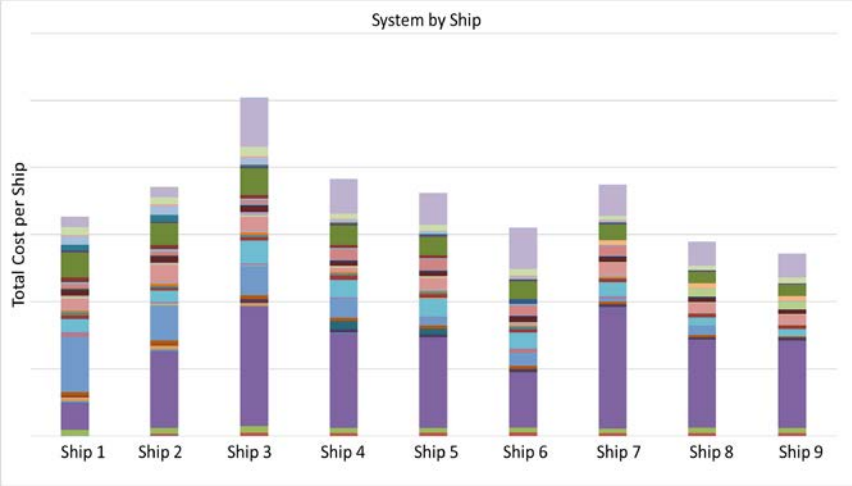

Figure 2: System by Ship

Comparing the different systems as in *Figure 2: System by Ship*, you can see that there are two systems that could be driving the total costs of Ship 3. The analyst should look closer at the large periwinkle system on the top of the bar and the larger purple system near the bottom of the bar. The periwinkle system at the top bar indicates the system experienced a cost increase that is being carried out to the other ships. There could be similarities between Ship 3 and Ship 7 for the lower purple system.

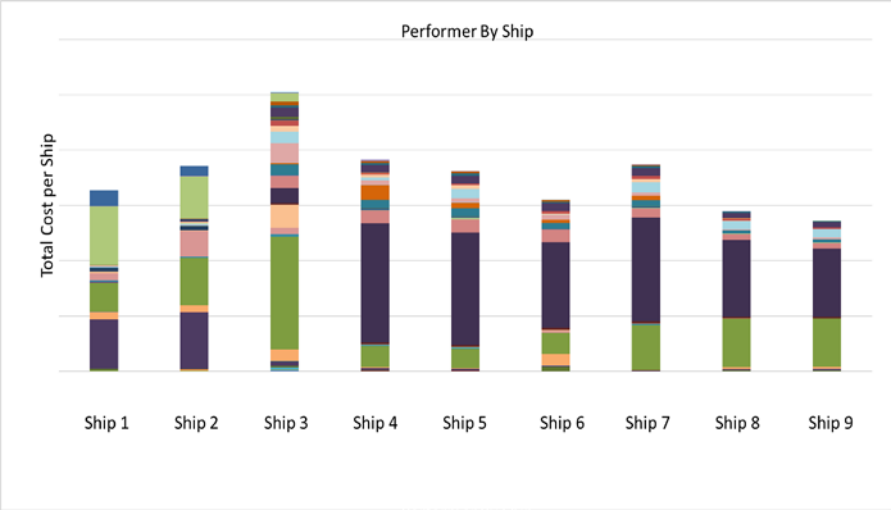

Figure 3: Performer by Ship

When we are viewing the *Figure 3: Performer by Ship*, we can see that there is a noticeable shift of costs among performers and the analyst can dig further into the ship systems and see how these changes are impacting ship systems.

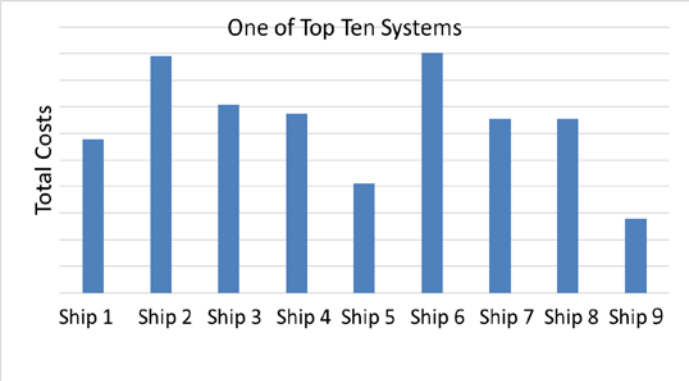

Figure 4: One of Top Ten Systems

Above in *Figure 4: One of Top Ten Systems*, indicates that the system cost per ship varies significantly from ship to ship. This give the analyst challenges to delve deeper into the requirements to determine if the system being estimated is more like one ship or another and thereby adjusting accordingly.

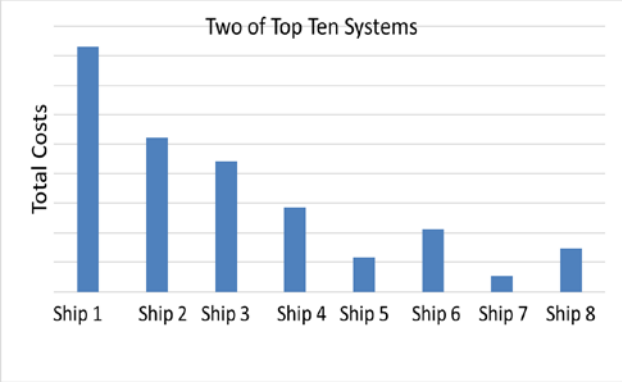

Figure 5: Two of Top Ten Systems

Occasionally you will see a system that will experience a decrease in costs at we see in *Figure 5: Two of Top Ten Systems*, above. The analyst need to determine if there is a learning curve impact with this system or that there is something peculiar with the system that is not found in other systems.

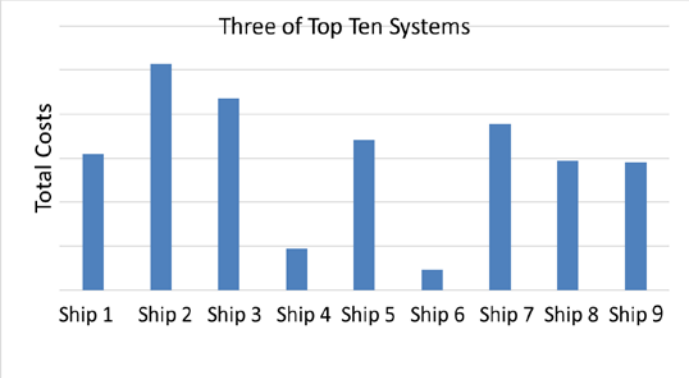

Figure 6: Three of the Top Ten Systems

When the analysts observe a system that is significantly lower for two of the nine ships as we see in *Figure 6: Three of the Top Ten Systems*, the analyst must inquire if the system installed on Ship 4 and Ship 6 are significantly unique or if there is an issue that is behind these ships that support the significant differences.

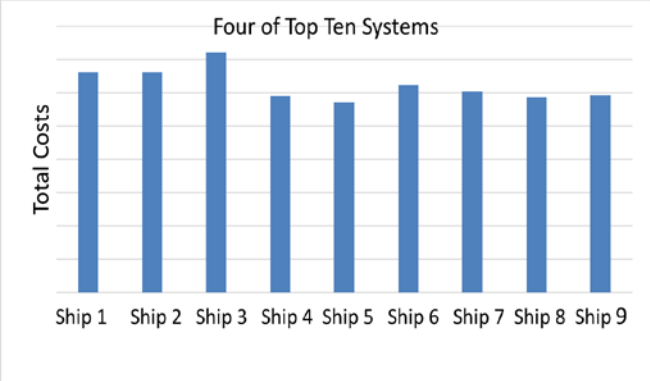

Figure 7: Four of Top Ten System

Many analysts expect the costs for a system to be as neatly described as in the system in *Figure 7: Four of Top Ten System*. The system cost is relatively stable and can be described in a seemingly high confidence cost element relationship (CER). The analyst will still need to do the math to verify if the CER is significant to use in a cost model.

General Application

This case was specific to us, but general methodologies and rules used here can be applied to many different situations. The analyst must first understand the data structure presented and then determine the best approach to glean the information that is buried in the database.

Many analysts have been given large unfamiliar raw databases to work with and analyze. This is common and many analysts fail to realize familiarities with other situations. When the data is reformatted, it becomes more useful and meaningful.

Highly structured and organized data may have a great appearance but is much less useable then if the data is not in a database format. Database format allows analysts to use excel functions and pivot tables however the analyst chooses to glean information from the data that can be applied to the situation.

There are endless possibilities and data views of a well formatted database, but not true of highly structured data.

Closing

This analysis and comparison of the different ship systems provided the foundation for which claims made in a Government Furnished Equipment (GFE) model could be tested. Such as comparisons between different ships for the same weapon system.

The transformation of this data allowed analysts to garner valuable insights into subtle inconsistencies and improve greater credibility and accuracy of the estimate.

If you're ever on the sending end of large data sets, be cognizant of the format it is received and the format you have it transformed. Structured data format tells one story, but not all of them, and may not be the most beneficial to the user.

Commented [SS1]: Not sure what you want to say here, I would reword this

Bibliography

GAO. (2009). Cost Estimating and Assessment Guide. In G. A. Office, *GAO Cost Etimating and Assessment Guilde, Chapter 10 Data.* Washington, DC: United States Government.

ICEAA. (2013). Cost Estimating Body of Knowledge. In I. C. (ICEAA), *CEBOK Module 4 Data Collection.* Washington, DC: ICEAA.