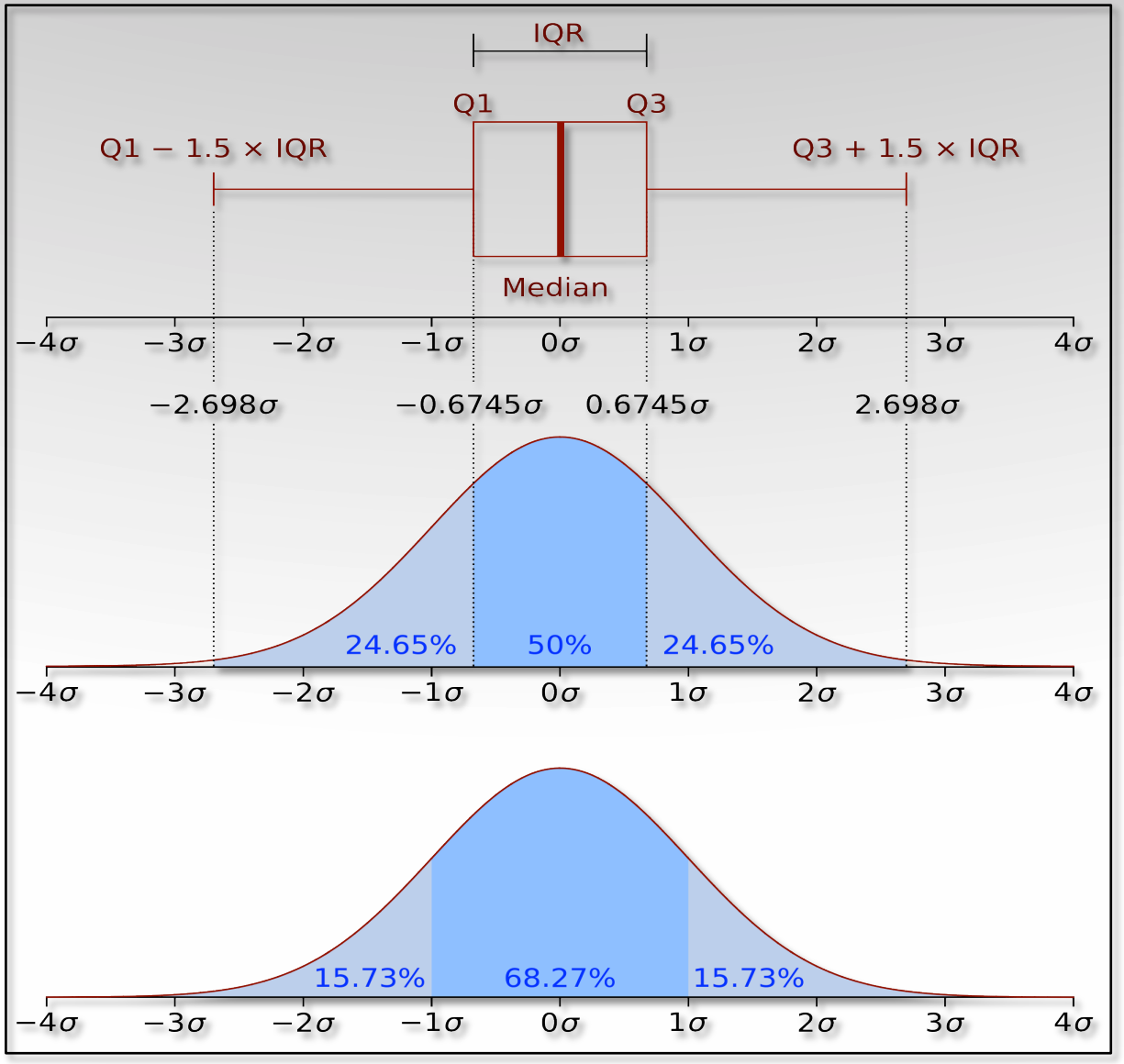
Mining Geology HQ Guidebook

Introduction to Exploratory Data Analysis (EDA) using Excel®



Erik C. Ronald

MINING GEOLOGY HQ

Mining Geology HQ Guidebook
Introduction to Exploratory Data Analysis (EDA) using Excel®

Erik C. Ronald, PG

1st Edition, 2016

Now that the legal stuff is out of the way, I hope you enjoy our guidebook and find it helpful in understanding and analyzing your geoscientific data.

Contents

Introduction

This guidebook is an attempt to provide a practical and easy-to-understand guide to exploratory data analysis (EDA) using software that is common in today's workplace. EDA is defined as an approach for analyzing a population of data to understand data characteristics, commonly with visual methods such as graphs. The target audience for this guidebook is industry geoscientists that are tasked with collecting, analyzing, and interpreting geoscientific datasets. Whether you're an exploration geologist looking at trace chemistry assays, a mine geologist concerned with production reconciliation, or a hydrogeologist trying to make sense of contaminants, data is data and it must be thoroughly analyzed to be understood prior to completing interpretation or modelling.

This guidebook outlines three types of EDA including: 1) univariate EDA concerned with the analysis of a single variable, 2) spatial EDA focused on the location and direction of data trends, and 3) bivariate EDA which analyzes the relationship between two variables.

The workflow follows a basic series of steps required by geoscientists to interrogate and analyze data. Most examples provided use geochemical assay data familiar to mining geoscientists such as drilling or point sample data. Only fundamental EDA is covered to establish a minimum baseline of data analysis. This guidebook does not discuss more sophisticated statistical tests or geostatistical analyses.

The reader will get the most out of this guidebook if you have at minimum some familiarity with Excel® and the basic concepts of spreadsheets and charts. Though this guidebook is meant as a step-by-step instructional guide, it does assume basic knowledge of computer use and rudimentary table and chart editing skills in Excel®.

Readers can perform all the EDA presented in this guidebook using only Excel® and the Data Analysis add-in that is provided with the software. No additional software is required to successfully utilize this guidebook. Additionally, there are many high quality statistical software packages available on the market and it is recommended that these various software be evaluated by the reader should more detailed analyses be desired.

The examples and step-by-step instructions are based on recent versions of Excel® for Windows PC. The author realizes that readers may have different versions resulting in slight differences between the guidebook's figures and what is displayed on the reader's computer. The fundamentals of EDA remain the same and the concepts are still applicable. The author apologizes for any confusion this may cause but please realize it is not practical to update our guidebooks every time a new software version is released.

A Note on Data Quality

Prior to performing EDA, the geoscientist must have confidence in the reliability and quality of the raw data. This guidebook does not aim to cover the subject of sampling theory, quality assurance and quality control (QA/QC), or the practice of database management. The old adage "garbage in – garbage out" holds especially true for data, data analysis, and geological interpretations. All too often geoscientists must deal with historic data of questionable quality due to myriad reasons. There is no point is wasting one's time with complicated estimation methods, detailed interpretations, or attempting to make informed decisions from fundamentally flawed data.

The guidebook assumes that by the time the reader is considering performing EDA, the data has already passed the fundamental checks for quality and applicability. If it has not, it is suggested to not bother with EDA but focus time and efforts on first obtaining a reliable and validated dataset. Only when you "tick this box" will the outputs and inferences learned through performing EDA be of any value.

I hope you enjoy this guidebook and find it useful in analyzing and understanding your geoscience data. For additional information on EDA and other geology-related topics, be sure to visit www.mininggeologyhq.com for articles and references.

Sincerely,

-Erik C. Ronald, PG

"He [the geologist] uses statistics as a drunken man uses lamp posts, for support rather than for illumination. "

- Andrew Lang (1844 – 1912), with minor modification, of course.

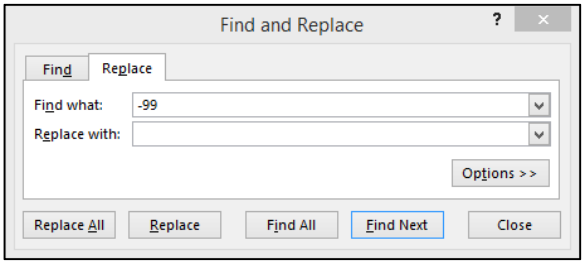
1 Null Values and Data Gaps

The first step of performing EDA on geoscientific data is to ensure the data is suitable for correctly calculating univariate (single variable) and bivariate (comparing two variables to each other) statistics. This initial step can also be thought of as data "cleaning".

a) *Identify which values are truly zero.* The difference between zero and null is an important distinction and has drastic implications on statistical outputs. A value of zero means you've measured a sample and the result was 0.00 (this is highly unlikely) while a null value simply means no data is available for that particular sample or interval. Spreadsheets may show a blank or a 0 so it's critical to know and understand the difference. In Excel®, zero means zero and blank means null.

It is common practice in database management to replace a blank cell with a designated null value. In geology specifically, it is standard for a null values to be represented as: "-99", "-9", or "-1". The reader can easily imagine what a few "-99" values can have on sample statistics so it is important to convert all null values to a blank cell prior to performing EDA in Excel®.

TIP: Use **Find>Replace** (Ctrl+F > Replace tab) to change all "null" values to a blank cell.



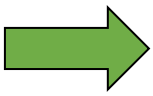
b) Ensure you're not *mixing numeric and character values*. There may be character values used when reporting raw data values from a laboratory that are below the equipment's ability to detect such as "nil", "n/a", "<mdl" or "bdl". Occasionally, this information is directly entered into a database resulting in the mixing of character and numeric fields. Some fields are obvious as to their meaning but others are not. With a bit of luck, there is a translation table available. Either way, EDA requires numeric values to calculate statistical values. All character entries need to be modified to be either a null value (blank cell) or modified accordingly to be numeric.

c) Determine and document how you handle *values below or above detection limit*. When a sample result is listed below laboratory testing limits, often the lab will provide a code stating this fact. As was mentioned in part b) above, values such as "<mdl", "trace", and "bdl" commonly all translate to "below the laboratory detection limit". There are a few methods of handling these data so whichever is chosen, supporting documentation must outline the "what and why". In lower detection cases, it is acceptable practice to replace the "bdl" character with a numeric value equal to half the detection limit. Therefore, for an element with a detection limit of 0.05, you'd replace all "bdl" values with 0.025. This will introduce bias to your data and many pure statisticians will not agree but as a geoscientist, you're likely more concerned with knowing the value is very low but not

zero and simply live with the introduced bias. In some cases, a value may be returned as "above detection limit". Though not as common, careful consideration must be taken to ensure these samples are flagged for re-testing and the final decision on how to manage the data is clearly documented.

d) *Data gaps* are common when dealing with drilling programs where samples are collected on regular intervals down hole using "FROM" and "TO" values. Missing intervals aren't detrimental (as long as they're not replaced with 0.00) though some geoscientists insert null values into those intervals to recognize a missing interval. This is important for good data management as the insertion of a null value clearly indicates a sample was not collected or tested. Beyond that, refer to part a) above and ensure all null values are replaced with a blank cell prior to performing EDA.

FROM	TO	AU
0	2	0.004
2	4	0.025
4	6	0.314
8	10	0.071



FROM	TO	AU
0	2	0.004
2	4	0.025
4	6	0.314
6	8	-99
8	10	0.071

e) *Significant figures*. It should be noted that significant figures can be a concern when data has been "well-traveled" through various software packages. In some cases, either due to deliberate reduction or software rounding, the significant figures of data values are erroneous. This can easily give a false sense of perceived accuracy or cause biases in the data when historic values have significantly different detection limits due to different generations of testing, different testing methods, or different laboratories.
It's always advisable to review the raw laboratory or original source data to understand the appropriate level of data accuracy. For instance, dealing with major oxide X-Ray Fluorescence (XRF) data which is precise to the nearest 0.1 percent in a dataset combined with Inductively Coupled Plasma (ICP) data for the same element that is accurate to the nearest 0.1 ppm can bias many output statistics.

f) *Overlapping intervals in drilling data*. It can be a common error in drilling datasets for FROM and TO intervals to overlap. The cause is typically due to the "fat finger" of human error when entering data but these issues will result in problems if attempting to make interpretations, inferences, or simply enter the data into mine planning software.

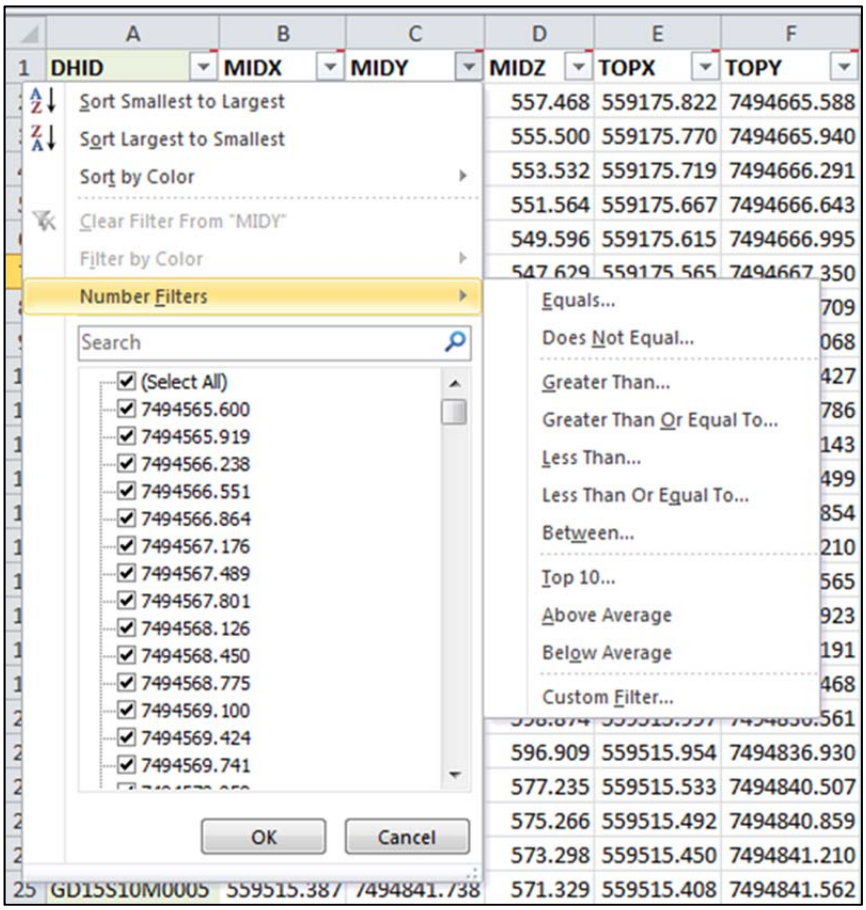
Overlapping intervals can be a challenge to identify in large datasets. In the case of drilling data, ensure a LENGTH field is entered which is simply the "TO" minus the "FROM" value. Performing EDA on this LENGTH field can, in many instances identify overlaps. Another method is to simply sort the FROM data from smallest to largest per drill hole and chart the data on a scatterplot. There are likely a dozen potential solutions for an issue like this that can be investigated further by using the steps and tools introduced in this guidebook.

2 Sort and Filter

The Sort & Filter function in Excel® is a powerful means of interrogating data which is accessed from **Home > Sort & Filter > Filter** 

(or Ctrl + Shift + L). It's one of the quickest ways to find those pesky negative or null values, along with errors in your data range. When you use this feature on categorical data, it's easy to find those misspellings from 1979 or the creative codes your summer student decided to add to the database.

1) Arrange the data into columns with headers in the first row.
2) Select the entire worksheet using Ctrl+A or the corner box between row 1 and column A, then click the **Home > Sort & Filter > Filter** button. The result will be that each column will have a small dropdown arrow next to it.
3) Click on the dropdown arrow. All the data is shown with a tick box to show or hide the data.



4) Scroll through the data or use the **Number Filters** option to be a bit more creative. The latter is recommended for large datasets.

! A word of caution when using **Sort & Filter:** Be careful when using filters on multiple columns at once as this hides data and can easily result in a confusing mess. It's recommended to keep things simple on the initial pass of data inquiry. Once you've interrogated individual variables then explore data groupings by adding your own columns and creating a Pivot table. This is ideal for interrogating data by geology, rock type, zone, or other designation.

3 Analysis ToolPak

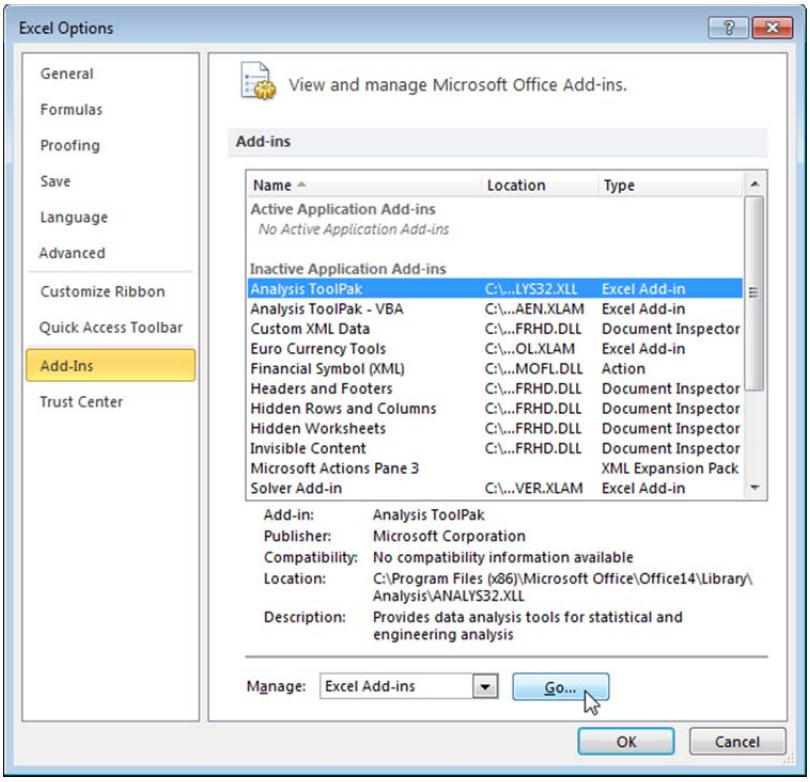
Excel® comes with a nice little add-in called the *Analysis TookPak*. Excel® 2016 and possibly future versions appear to be incorporating more statistical analyses options in the standard loaded software including the ability to calculate histograms directly from the Chart options. As this latest version of Excel® is being released in the same year as this guidebook, it is assumed the reader is working with an earlier version thus requiring the add-in.

The *Analysis Toolpak* add-in does not come automatically loaded so you'll need to activate it as follows:

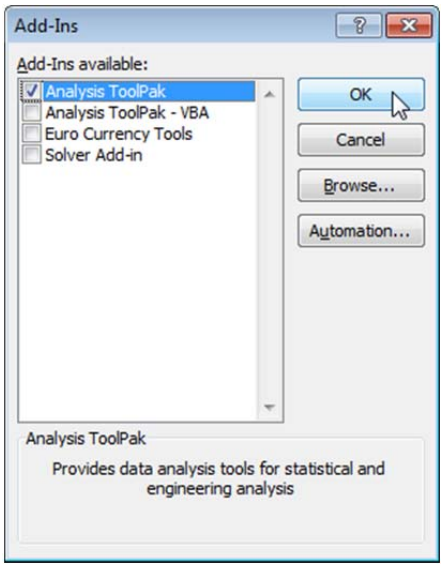
1) Click on the **File** > **Options** (located at the bottom of the menu)

   In older versions, you may need to use the **Microsoft Office Button**  .

2) Click **Add-Ins**, then in the **Manage** field at the bottom, select **Excel Add-ins**.

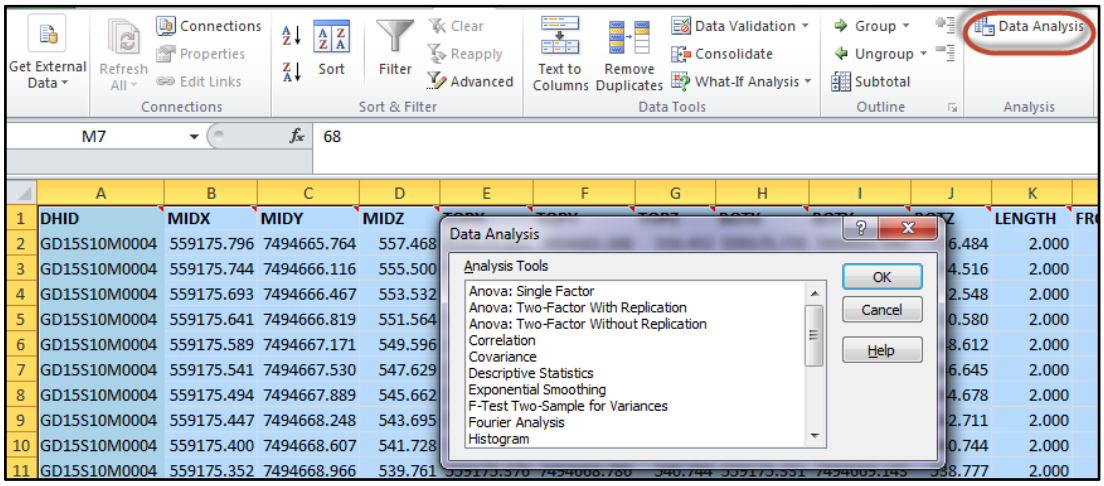


3) Click **Go**.
4) In the Add-Ins available box, select the **Analysis ToolPak** check box, and then click **OK**.



TIP: If the *Analysis ToolPak* is not listed in the Add-Ins available box, click Browse to locate it on your internal hard drive under **Programs**.

Once the *Analysis ToolPak* is loaded, the **Data Analysis** command is available in the **DATA** > **Analysis** portion at the top of the screen (look to the far right).



Many of the proceeding sections assume you've successfully loaded the *Analysis Toolpak*. Only the more commonly used menu items in the ToolPak are described in this Guidebook. It is always encouraged to explore other menu items and utilize the Help function or simply perform an internet search on terms of interest to gain additional background and understanding as to whether other statistical tests will be beneficial for your particular dataset.

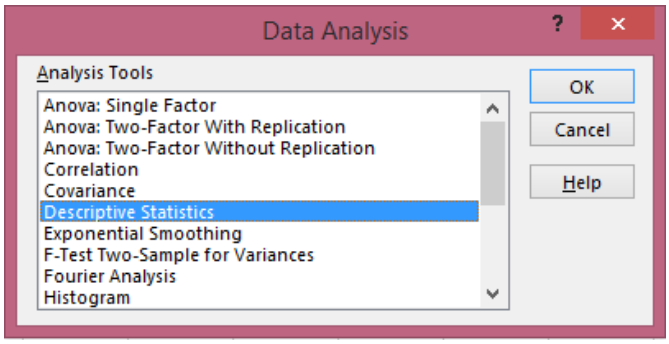
TIP: There are third-party Excel® add-ins available to perform various statistical data analyses. Each has its merits and should be investigated further if you plan on performing more rigorous data analysis using Excel® in the future. There are a variety of hyperlinks available to Excel®-based statistical add-ins available at www.mininggeologyhq.com/resource-geology

Additionally, most professional geoscientists have access to geological modeling, mapping, and planning tools that likely include a statistical analysis package. It is up to the individual as to their preference and needs but the *Analysis Toolpak* does provide an excellent option at no additional cost along with the ability to output familiar charts and tables.

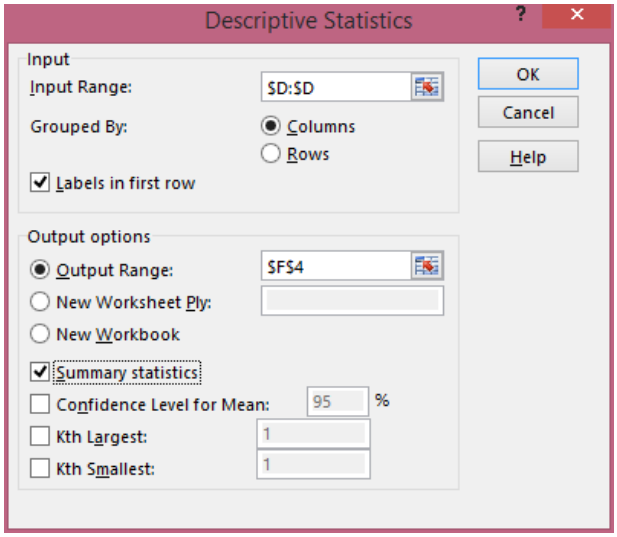
4 Descriptive Statistics

The first three steps in this guidebook have prepared us for performing EDA. Now the first type of EDA covered is calculating univariate statistics. The **Descriptive Statistics** option in the *Analysis Toolpak* is an easy way of calculating multiple statistical properties such as the mean, maximum, minimum, and other descriptors quickly for either a single or for multiple variables.

1. Arrange data in columns by variable.
2. Click **DATA** > **Data Analysis**
3. Select **Descriptive Statistics** from the menu then click **OK**.



4. Select the data by using the Input Range. This can be a single variable (as shown below where column D includes all the data of interest) or select multiple variables at once. When multiple variables are selected, **Descriptive Statistics** will calculate univariate statistics for each variable. Performing bivariate or multivariate analyses is addressed in later sections.
5. Tick the **Labels in first row** box if you're including the data header or label.
6. Next, under **Output options** select where you'd like the summary statistical outputs to go – same page, a new tab or an entirely new workbook. The example below will place the output table in the same worksheet starting at cell F4.
7. Tick the **Summary statistics** box. Click **OK**.



The output will be two columns (when selecting a single variable): the statistical property and the corresponding value. It is recommended to adjust your significant figures as the default output returns an unreasonably large amount of figures for some items.

Column1	
Mean	54.89
Standard Error	2.874404
Median	59
Mode	67
Standard Deviation	28.74404
Sample Variance	826.2201
Kurtosis	-1.12946
Skewness	-0.24254
Range	99
Minimum	1
Maximum	100
Sum	5489
Count	100

Mean is the arithmetic average and a value most people are familiar with when considering the average value of a data population.

Standard Error (SE) is calculated by dividing the standard deviation (δ) by the square root of the total number of samples (count). The SE can indicate how close the sample mean is from the "true" population when you consider your data represents merely a small sample of a greater or "true" population. In other words, if you have a small number of samples and a high SE, it's likely your mean would be different if you doubled or tripled your sample count. If your SE is small, it can be thought of as having higher confidence in your mean even if you collect more samples from that population. To use an example, if you collected just four samples from a highly variable conglomerate for grain size, chances are your mean grain size isn't that representative of the entire Formation or unit. Therefore, in this particular case the SE would be high relative to the mean.

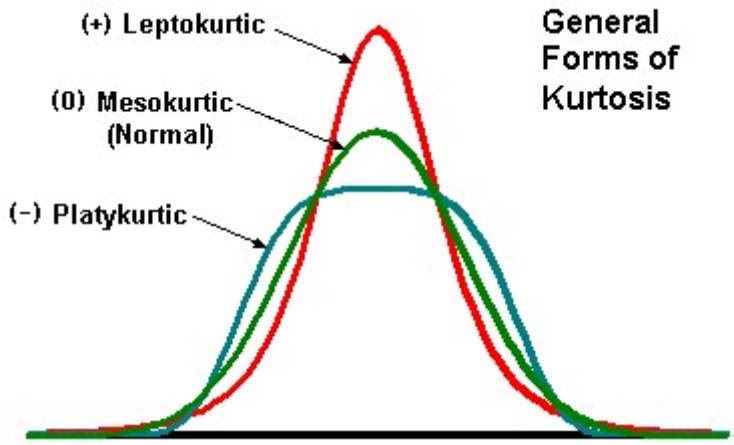
The **Median** (50th percentile, cumulative frequency = 0.5, or middle quartile) is the value half way from the minimum and maximum values in the population. The median tends to be more stable as it is not as susceptible to extreme outliers as the mean. In a case where there is a large difference between the Mean and Median it's a good indication of the presence of outliers in the data population.

The **Mode** is the most common number in your dataset. It can also be thought of as the number having the highest likelihood of occurring in your data. Typically in geoscientific data, the mode isn't terribly valuable as most geoscientists are concerned with average values or the average above a particular cut-off or threshold.

Sample Standard Deviation (δ) and **Variance** (δ^2) are measures of data dispersion. They measure the spread of data across the sample distribution. The main differences are the units. Standard deviation is expressed in the original data units thus most people find this easier to understand than a square of units. Variance is more commonly used in geostatistical calculations.

Kurtosis, though sounding like a bad disease, it is a measure of the shape of the distribution or the peakedness of the curve. Negative kurtosis can be an indicator of the presence of two populations of data with different means while positive kurtosis can indicate two populations

with different standard deviations. In many cases, high kurtosis is the result of very low variability of a sub-population or grouping of data.

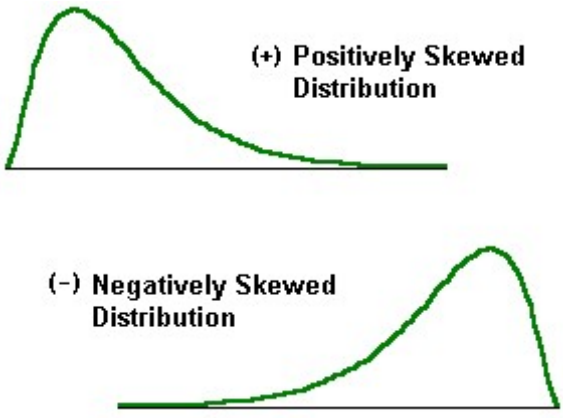


Skewness provides an indication of the data distribution shape. Typically this is easier to understand once you can look at a graphical output of a histogram or frequency distribution. The **Coefficient of Skewness** or asymmetry is a measure of the shape of the distribution and tail:

Positive value (+) means it has a long tail of large values.

Negative value (-) means it has a long tail of small values.

Zero value means the distribution is perfectly symmetrical (mean $\simeq$ median).



A positively skewed distribution is common in geochemistry where the long "tail" may represent outlier data of rare high-concentrations of trace elements (such as As or Au) while the majority of the distribution is diluted or dispersed in the system. The opposite may occur with negatively skewed data but for major elements or oxides. For example, SiO_2 in a granitic intrusive system would be negatively skewed.

Range, Minimum, and **Maximum** are obvious as to what they represent. It's always a good idea to check your data limits for what makes sense as this is usually one of the first indications

of the presence of assay values below the detection limit, errors, or coordinates in the middle of the Pacific.

The **Count** and **Sum** hopefully don't require further explanation. The count is important to note especially when dealing with grouped or domained data sets. In many cases when too many data sub-divisions are used, such as in stratigraphy, the count per division becomes small thus resulting in high SE and lowering the confidence in the statistics. The Sum is rarely used in most cases for geology-related data.

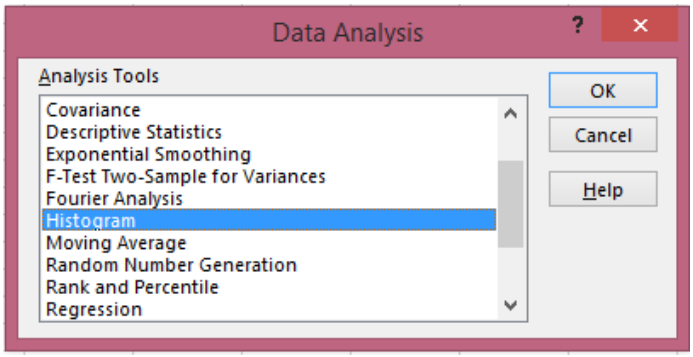
Though not included in Excel® Descriptive Statistics, it's recommended to quickly calculate the **Coefficient of Variation** (CV). The CV is expressed as a percentage and is equal to: CV = standard deviation (δ) / mean (m). As a rule of thumb, data populations with CV greater than 1 tend to comprise multiple populations or contain outlier data. CV is a nice check as to whether more detailed outlier analysis will be required or the appropriateness of sub-grouping of data.

5 Histogram and Frequency Distribution

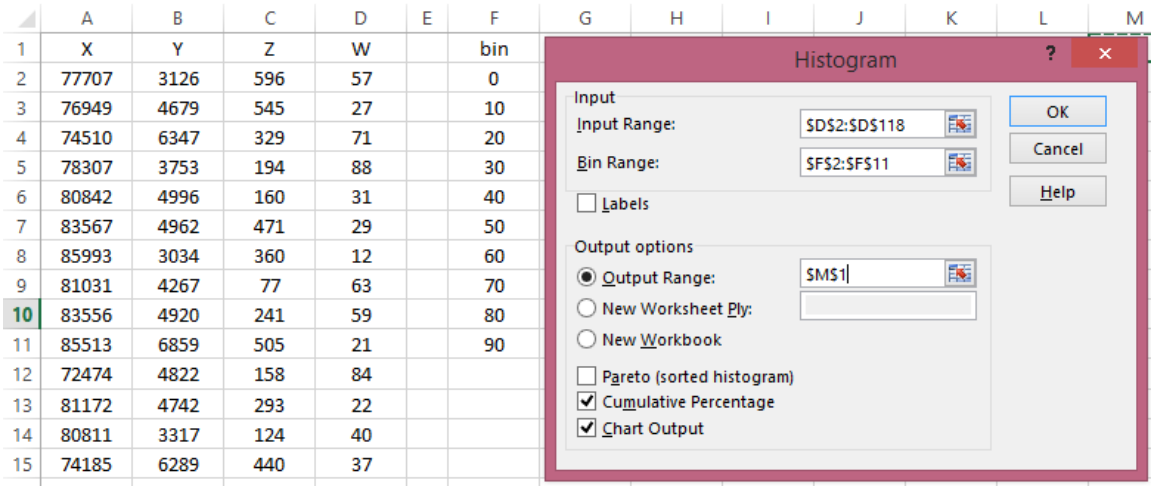
Histograms graphically represent the distribution of the numeric data population. Prior to creating a histogram (or frequency distribution diagram) you'll need to have all the data for a particular variable of interest in a single column and the bins in a second column (see figure in step 3 below). A **bin** or class interval is how your data will be divided into groups Bins must be consecutive and non-overlapping intervals of the variable but do not need to be the same size. It's helpful to review the Descriptive Statistics to determine the bin size based on the Range, Standard Deviation, or a particular set of parameters you have such as cut-off grades or thresholds.

TIP: A rule of thumb for bin size is to divide the range by roughly half the standard deviation. Alternatively, many statisticians simply chose 15-30 bins from the minimum to the maximum values then adjust accordingly.

1. Determine the size, number, and cut-off for your intervals or bins. Enter these in a column (as shown below under Column F: *bin*).
2. Select **DATA** > **Data Analyzer**, then select **Histogram** from the menu. Click **OK**.



3. Choose the **Input Range** for the data (ex. column D) and the **Bin Range** (ex. column F).

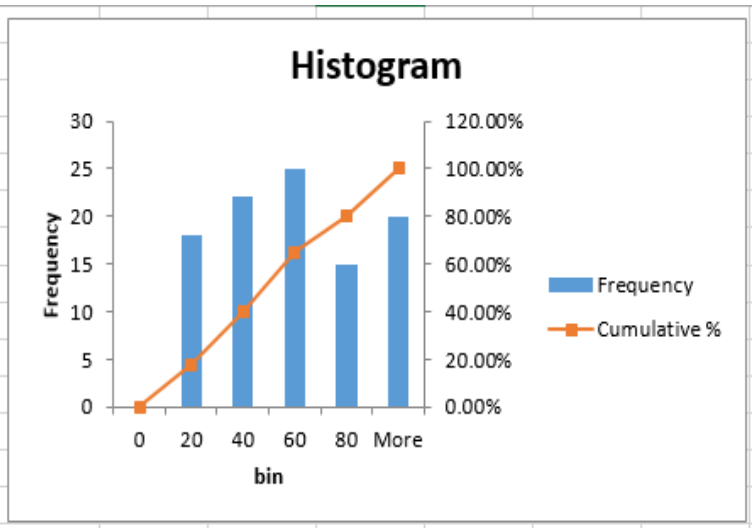


	A	B	C	D	E	F
1	X	Y	Z	W		bin
2	77707	3126	596	57		0
3	76949	4679	545	27		10
4	74510	6347	329	71		20
5	78307	3753	194	88		30
6	80842	4996	160	31		40
7	83567	4962	471	29		50
8	85993	3034	360	12		60
9	81031	4267	77	63		70
10	83556	4920	241	59		80
11	85513	6859	505	21		90
12	72474	4822	158	84		
13	81172	4742	293	22		
14	80811	3317	124	40		
15	74185	6289	440	37		

4. Choose where the table and chart will be placed in **Output options**.
5. Be sure to select **Chart Output** or you'll simply get a table of values.

6. Additionally, it may be helpful to click the **Cumulative Percentage** box as another means of visualizing your data. Click **OK**.
7. The output will be a table and graph as below:

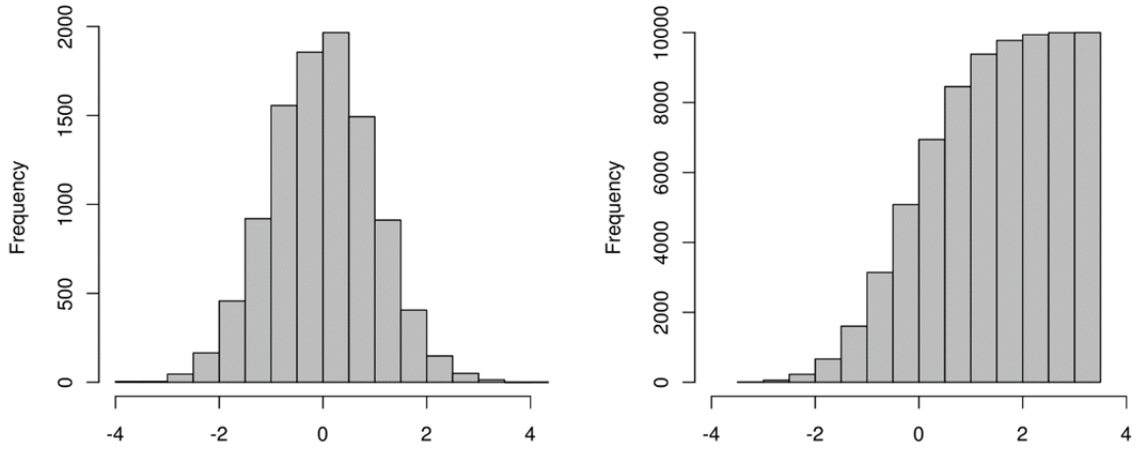
bin	Frequency	Cumulative %
0	0	0.00%
20	18	18.00%
40	22	40.00%
60	25	65.00%
80	15	80.00%
More	20	100.00%



8. Now modify the legend, title, axes and other properties of the histogram chart to suit your needs.
9. If you'd like to remove the space between the bars, right click on a bar, select **Format Data Series** and change the **Gap Width** to 0%.

It's advisable to modify your bins so you can gain an appropriate insight into the data distribution. If the bins are too large or too small, the distribution may be difficult to interpret. It is common practice to use bins of equal size but not necessary. When determining the frequency above a threshold or cut-off you may have all values under the threshold in a single bin.

The figure below shows a normal distribution represented as a histogram (left) and a cumulate percentage distribution (right). These are the same distribution but merely presented in different ways. Additionally, histograms are often displayed as a curve instead of a column chart. Ultimately it is up to you to decide the preferred way of visualizing the data.

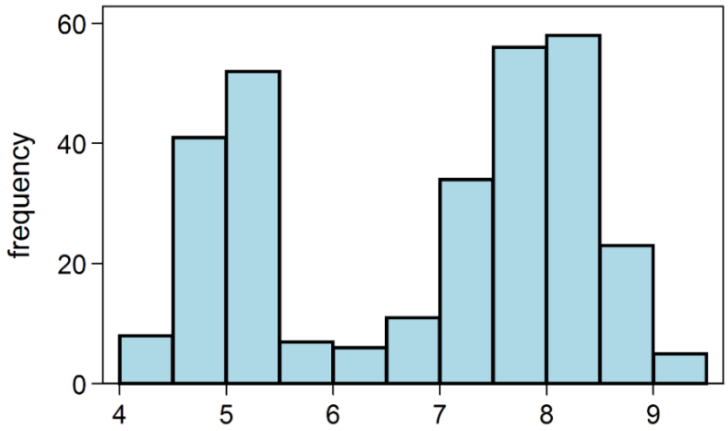


There are three main properties of a histogram to consider and interpret:

1) **Shape** – the symmetry/asymmetry of the frequency distribution;
2) **Position** – where the "average" position of the whole distribution is along the scale;
3) **Dispersion** – how spread-out the distribution is along the scale.

If you recall from *Descriptive Statistics*, we have already calculated the majority of statistical measures, but viewing the data as a histogram will allow for a better understanding of the data's behavior and characteristics. For instance, the Mean, Mode, and Median all inform about the position of the distribution. Dispersion is the degree of variability or how spread-out the data is along the x-axis. We have already calculated three measures of dispersion in *Descriptive Statistics*. These being the standard deviation, variance, and range. Lastly, the shape is represented by the skewness and kurtosis.

In some cases, the distribution may exhibit polymodal distribution such that there is more than one "hump" in the data. Commonly in geoscience, data contains two peaks thus we use the term bimodal distribution. The presence of a polymodal distribution can be an indicator of your data containing two or more distinct populations with different means that should likely be separated and analyzed individually.



A histogram exhibiting a bimodal distribution.

6 Quantiles

It can be informative to divide your sample population into equal parts or intervals. These divisions are generically termed quantiles. Common divisions include 100 parts called percentiles, ten parts called deciles, and four parts called quartiles. The most common are percentiles (P) and quartiles (Q). Therefore the first quartile represented by Q1 is equal to the 25th percentile or P25.

1) Arrange your data into a single column.
2) In an adjacent cell, type in the =PERCENTILE function. This function requires a range of data followed by the percentile to calculate represented as a decimal from 0 to 1 (e.g. =PERCENTILE(A1:A526,0.3) will return the 30th percentile). This means that 30% of the data is lower than the value returned (ex. 29 as shown below).

	A	B	C
1	12		
2	64	=PERCENTILE(A1:A526,0.3)	29
3	18		
4	43		
5	18		
6	14		
7	39		
8	26		
9	1		
10	76		
11	87		
12	98		
13	54		
14	56		

3) Quartiles are calculated in a similar way. The formula is =QUARTILE(range,quartile). For example, =QUARTILE(A1:A526, 3) will return the Q3, third quartile, or 75th percentile. The second value in the formula must be an integer from 0 to 4 realizing that Q0 is the minimum, Q2 is the median, and Q4 is the maximum.

	A	B	C
1	12		
2	64	=PERCENTILE(A1:A526,0.3)	29
3	18		
4	43	=QUARTILE(A1:A526,3)	72
5	18		
6	14		
7	39		
8	26		
9	1		

To calculate deciles or other equal parts of the distribution, simply use the PERCENTILE function with the appropriate values such as 0.1, 0.2, 0.3 and so forth.

TIP: It should be recognized that as of Excel® 2013, the software has three methods of calculating percentiles and quartiles. Microsoft™ have stated that the classical PERCENTILE and QUARTILE functions used in previous versions should not be used. Instead it is

recommended to use the *.INC or *.EXC functions added with Excel® 2013 and available in newer versions.

If you need to repeat calculations from historic Excel® versions, the PERCENTILE and PERCENTILE.INC returns the same values. PERCENTILE.INC is an inclusive function whereby for any value in the percentile from zero to one (0 – 1) you will get a result. PERCENTILE.EXC will return an error if you use a value which is outside the range for the data set. When in doubt, use PERCENTILE.INC.

7 Interquartile Range and Outliers

One method of interrogating the variability of a data population is by using the interquartile range or IQR. As stated in the previous section, quartiles break the data distribution into four equal parts between these five values: Q0 (minimum value), Q1 (25th percentile), Q2 (median), Q3 (75th percentile) and Q4 (maximum value). The equation for calculating the **IQR = (Q3-Q1)**.

When a data population has outliers or is skewed, the total range may not be a good indicator of the variability where the IQR can be more helpful as it covers 50% of the population – from the 25th to 75th percentile. Another useful aspect of the IQR is the ability to calculate and identify outliers and extreme outliers. The following equations are used to determine the thresholds at which samples are considered outliers or extreme outliers:

Outliers:

Upper Outliers = Q3 + 1.5 *(IQR)
Lower Outliers = Q1 – 1.5 *(IQR)

Extreme Outliers:

Upper Extreme Outliers = Q3 + 3 *(IQR)
Lower Extreme Outliers = Q1 – 3 *(IQR)

The IQR and outlier definitions are straightforward to calculate in Excel®:

1) Arrange your data in a column (shown in column A below).
2) Use the QUARTILE.INC function as described in the previous section to calculate Q1 and Q3.
3) Type in the formula: Q3 – Q1 (=D3-D2 in example). The result is your **IQR**.

	A	B	C	D
1	12	*Item*	*forumula*	*value*
2	64	**Q1**	=QUARTILE.inc(A1:A101,1)	21
3	18	**Q3**	=QUARTILE.inc(A1:A101,3)	66
4	43	**IQR**	=Q3-Q1	45
5	18			
6	14	Upper Outlier	=Q3+1.5*(IQR)	133.5
7	39	Lower Outlier	=Q1-1.5*(IQR)	-46.5
8	26			
9	1	Ext Upp Outlier	=Q3+3*(IQR)	201
10	76	Ext Low Outlier	=Q1-3*(IQR)	-114
11	21			
12	65	Maximum	=max (A1:A101)	100
13	48	Minimum	= min(A1:A101)	1
14	2	Range	= maximum - minimum	99
15	87			

In the example shown above, column A includes 100 random numbers between 1 and 100 (only the first 15 are shown). The minimum value is 1 with the maximum of 100 (range = 99). In this case, our test for outlier and extreme outliers returned values beyond the range of our datasets. So no outliers are present in our trial data from column A.

In the case where outliers are present, the definition for outlier and extreme outlier provides a threshold. All values beyond the threshold are considered outliers. It's up to the geoscientist analyzing the data as to whether outliers are a concern or not but they should be identified at a minimum. In the next step, we'll review how to generate a Box and Whisker plot which produces a graphical output of the IQR and outlier bounds.

A Note on Outliers:

By definition, an outlier is any point that is distant or distinct from the rest of a data population. The reasons for the presence of an outlier vary greatly but the common causes are errors in the data (sampling, laboratory, or otherwise), poorly grouped data so there are actually two or more populations (contamination from another zone), or just highly skewed data (such as high-nugget Au values).

Some statisticians will remove outliers from a sample population prior to calculating statistical values such as the mean and variance while others take a traditional approach and include them. Whichever option is chosen matters less than ensuring the reasons are well documented. In mining and exploration of metal deposits, the industry is focused on finding and exploiting outliers and anomalous metal occurrences. Deposits of economic significance are by definition outliers. Industry professionals must realize that in low-concentration metal exploration, the outliers of the rare extreme high grade samples are of significant interest and typically contain the highest total metal per tonne and therefore should not be simply discarded due to statistical conformity. In short, caution should be taken before discarding, top-cutting, or reducing the influence of high-yield variables but the impact of the high magnitude samples on a general data population must always be clearly understood. Conversely, if extreme outlier deleterious variables occur within a high grade zone, the outlier values may skew the average deleterious variable to exceed a threshold therefore reducing the entire high-grade zone to waste. One must always be careful when dealing with highly skewed data and outliers and understand their effects on statistical outputs.

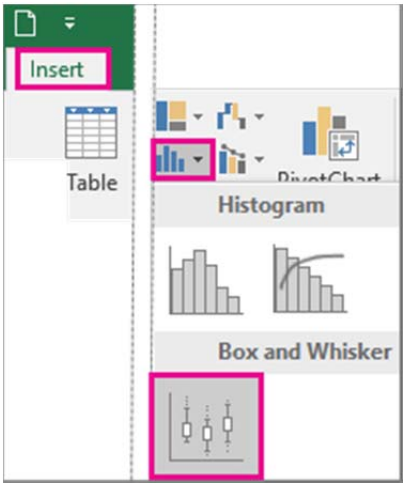
The presence of outlier data has significant implications on various estimation techniques which assume an underlying normal distribution (such as Ordinary Kriging). These estimation techniques are very sensitive to the presence of outliers data. Outliers can greatly bias the spatial continuity of a variable resulting in erroneous estimation. There are many methods of dealing with this in regard to estimation which are beyond the scope of this guidebook. At minimum, geoscientists must identify and document the presence and nature of outlier data and appreciate their impact on EDA.

8 Box and Whisker Plots

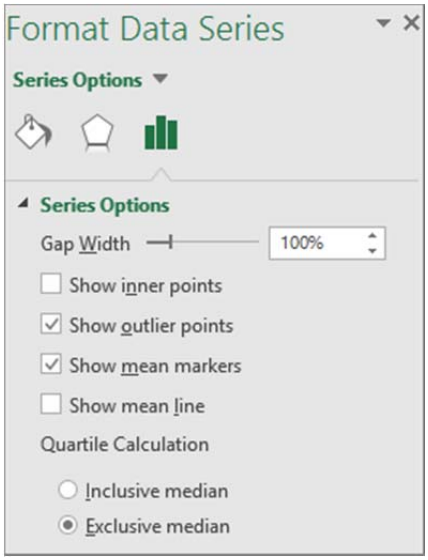
Box and Whisker plots are wonderful graphs that display a great deal of statistical information in one place. The rather silly name is in fact perfectly descriptive; the chart is literally made up of a box and a "whisker" on either side representing the outlier thresholds. These charts include the median, Q1, Q3, IQR, and outlier thresholds. The good news: if you are using Excel® 2016 then Box and Whisker plots are included as a ready-made chart. The bad news: if you using any version prior to 2016, it will be a lot of manual work.

In Excel® 2016:

1) Arrange the data into columns. Select the variable or variables of interest.
2) Click **Insert > Insert Statistic Chart > Box and Whisker**



3) To modify the chart options. Right click on a box to select then click **Format Data Series**.
4) In the **Format Data Series** pane, with **Series Options** selected, make the changes that you want.



5) The resultant graphic will be a Box and Whisker displaying mean, median, Q1, Q3 and outlier thresholds.

Excel® 2013 and previous versions:

This example is from a randomly generated dataset of 100 values. The process of graph creation is quite manual and time consuming but results in an acceptable "work around" to generate a Box and Whisker plot in Excel® 2013 or earlier version. It is not recommended to manually create Box and Whiskers if you have a lot of variables unless you don't mind spending hours manipulating charts.

1) Calculate the required statistics if you haven't already through *Descriptive Statistics*. This includes: Q0 (minimum), Q1, Q2 (median), Q3, Q4 (maximum), Lower Outlier (L.O.), and Upper Outlier (U.O.). In this example, the raw data is located in column A.

	A	B	C	D	E
1	12				
2	64				
3	18				
4	43				
5	18				
6	14				
7	39				
8	26				
9	1				

Item	*Value*	*Forumla*
Minimum	2	=MIN(A1:A101)
Lower outlier	8	=D5-(1.5*(D7-D5))
Q1	21	=QUARTILE.INC(A1:A101,1)
Median	48	=MEDIAN(A1:A101)
Q3	66	=QUARTILE.INC(A1:A101,3)
Upper Outlier	93	=D7+(1.5*(D7-D5))
Maximum	99	=MAX(A1:A101)

2) Next, calculate the differences between these values as shown in the table below to generate the chart output.

	A	B	C	D	E
1	12				
2	64				
3	18				
4	43				
5	18				
6	14				
7	39				
8	26				
9	1				
10	76				
11	21				
12	65				
13	48				
14	2				
15	87				
16	66				

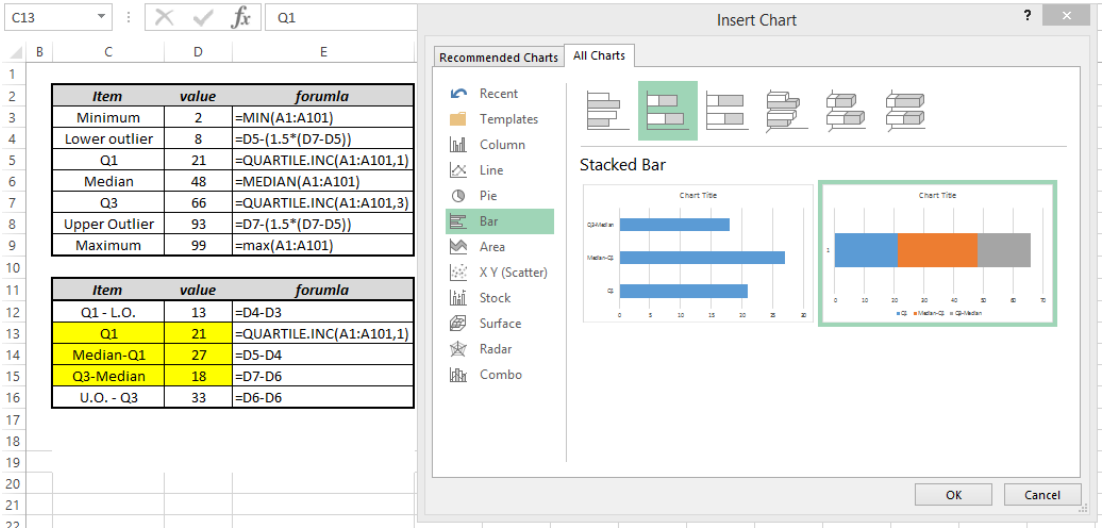
Item	*Value*	*Forumla*
Minimum	2	=MIN(A1:A101)
Lower outlier	8	=D5-(1.5*(D7-D5))
Q1	21	=QUARTILE.INC(A1:A101,1)
Median	48	=MEDIAN(A1:A101)
Q3	66	=QUARTILE.INC(A1:A101,3)
Upper Outlier	93	=D7+(1.5*(D7-D5))
Maximum	99	=MAX(A1:A101)

Item	*Value*	*Forumla*
Q1 - L.O.	13	=D5-D4
Q1	21	=QUARTILE.INC(A1:A101,1)
Median-Q1	27	=D6-D5
Q3-Median	18	=D7-D6
U.O. - Q3	33	=D8-D7

3) Highlight the cells as shown in yellow:

Item	*Value*	*Forumla*
Q1 - L.O.	13	=D5-D4
Q1	21	=QUARTILE.INC(A1:A101,1)
Median-Q1	27	=D6-D5
Q3-Median	18	=D7-D6
U.O. - Q3	33	=D8-D7

4) Go to **Insert > Chart**. Click the See All Charts button (click the little arrow in the lower right corner) and find the **All Charts > Bar > Stacked Bar** chart as shown:

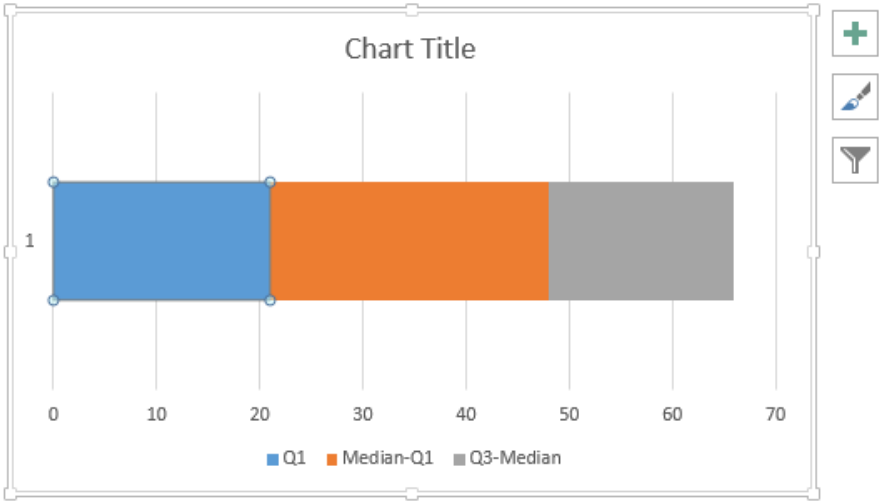


Item	value	forumla
Minimum	2	=MIN(A1:A101)
Lower outlier	8	=D5-(1.5*(D7-D5))
Q1	21	=QUARTILE.INC(A1:A101,1)
Median	48	=MEDIAN(A1:A101)
Q3	66	=QUARTILE.INC(A1:A101,3)
Upper Outlier	93	=D7-(1.5*(D7-D5))
Maximum	99	=max(A1:A101)

Item	value	forumla
Q1 - L.O.	13	=D4-D3
Q1	21	=QUARTILE.INC(A1:A101,1)
Median-Q1	27	=D5-D4
Q3-Median	18	=D7-D6
U.O. - Q3	33	=D6-D6

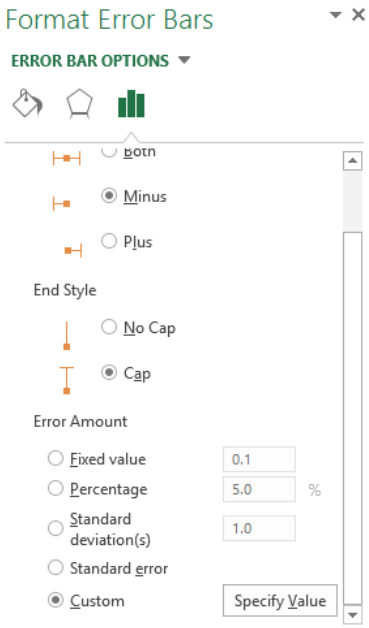
5) Click **OK**.

6) The next step involves adding the Lower Outlier and Upper Outlier "whiskers" to the chart. Select the Q1 box on the chart and click the **Chart Elements** button located on the right side of the chart.



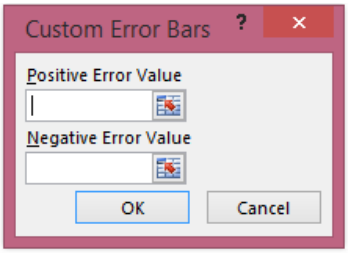
7) Check the **Error Bars** option and click the small arrow to the right > **More Options**.

8) This opens the **Format Error Bars** options. As the goal here is to enter the lower "whisker", we'll click the Minus toggle, End Style to be a Cap, and select the Error Amount to be Custom as shown below.



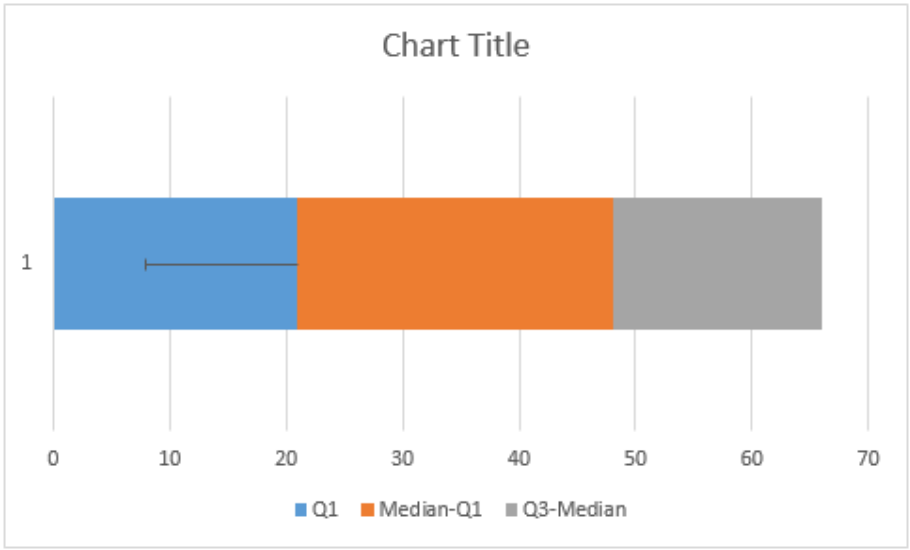
9) Click the **Specify Value** in the Custom field. Leave the Positive Error Value field as default. Modify the Negative Error Value to be equal to the Q1-L.O. value from the data table as shown below:

Item	*Value*	*Forumla*
Q1 - L.O.	13	=D5-D4
Q1	21	=QUARTILE.INC(A1:A101,1)
Median-Q1	27	=D6-D5
Q3-Median	18	=D7-D6
U.O. - Q3	33	=D8-D7



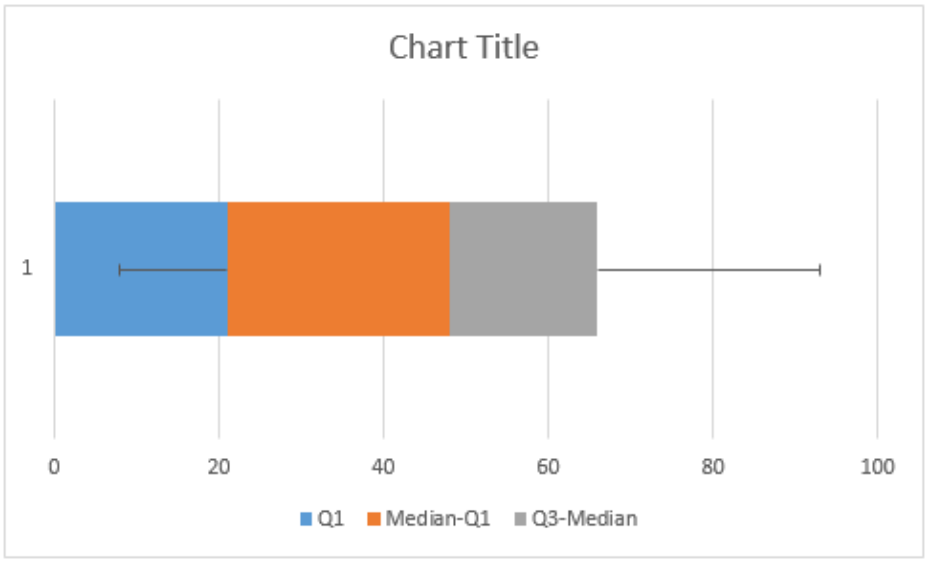
10) Click **OK**.

11) The resultant chart will resemble:

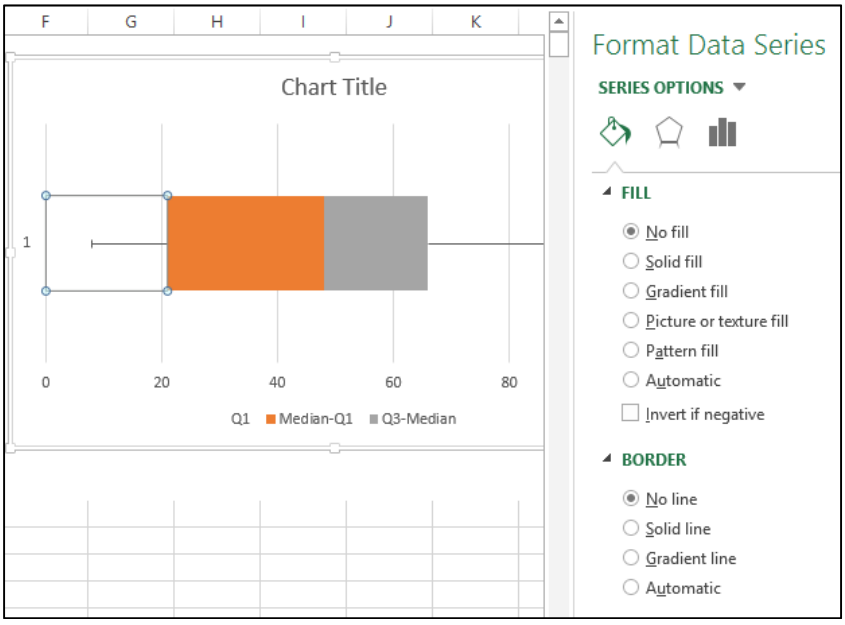


12) Now repeat the steps for the Q3-Median box (grey box in example) but keep in mind, this time we're concerned with the **Positive Error Value** that will equal the **U.O.-Q3** from our table.

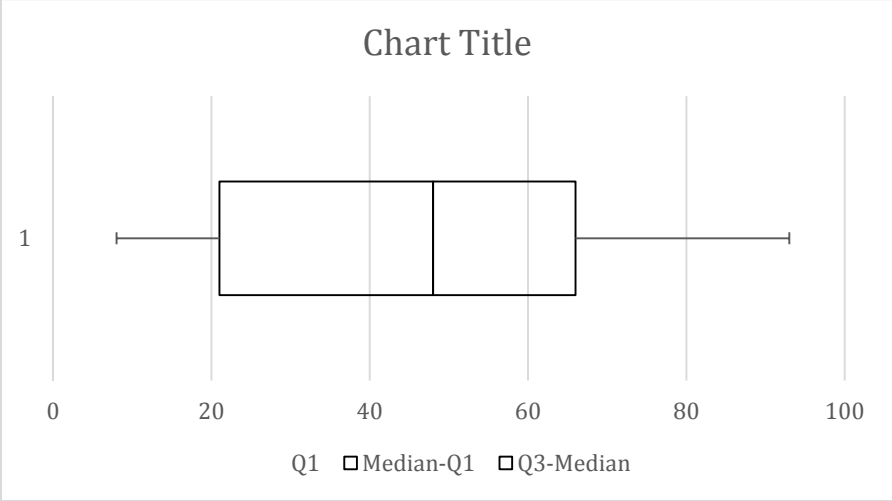
13) Upon completing that step, the chart should resemble:



14) The next steps are to remove the box fill and modify the borders for our three data points. To start, right click on the Q1 data box (blue area in example) > **Format Data Series**. Modify these options to be **No fill** and **No line** as shown:



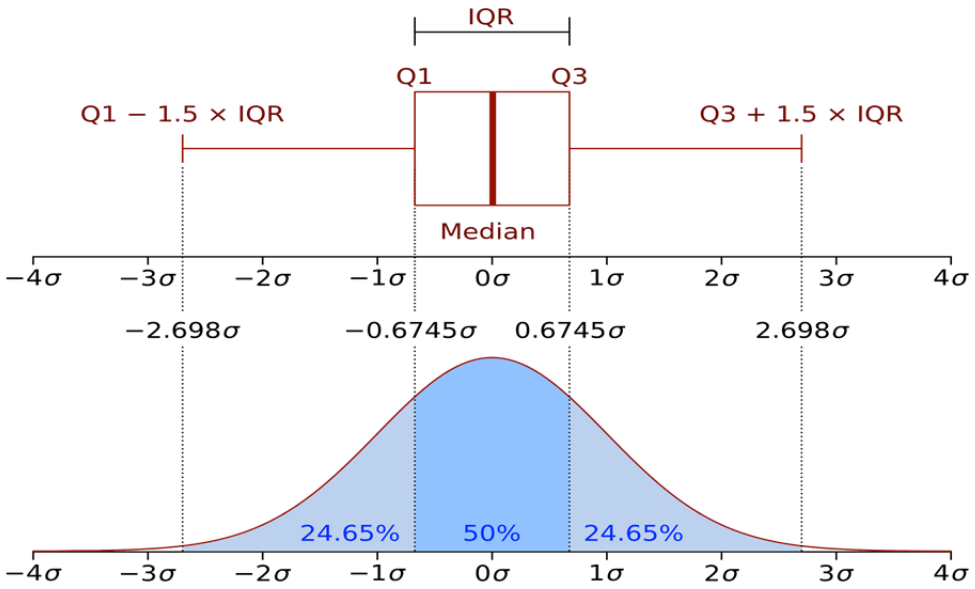
15) Next, modify the other two areas to be **Fill>No fill** and **Border>Solid line**. The resultant chart should resemble:



16) Lastly, it's good practice to modify your data title, Chart title, legend and any other characteristics of the chart to suit your needs.

As mentioned previously, this chart may not be considered by some to be a true Box and Whisker as the Mean is not represented and outlier samples are not shown on the graph (just the thresholds). Regardless, this chart can illustrate a great deal of information about the data distribution and is useful when comparing multiple variables side by side or showing a cut-off grade/threshold value visually compared with the median, Q1, Q3, and other values.

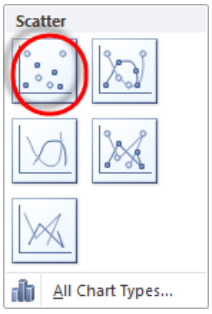
The figure below shows how similar statistical values for a data distribution are represented on both a Box and Whisker and a frequency distribution. The x-axis scale displays the number of standard deviations in the distribution. This figure is useful in visualizing a data population and the various ways of representing the same data and understanding the relationships between the various statistical parameters.



9 Spatial Data Maps

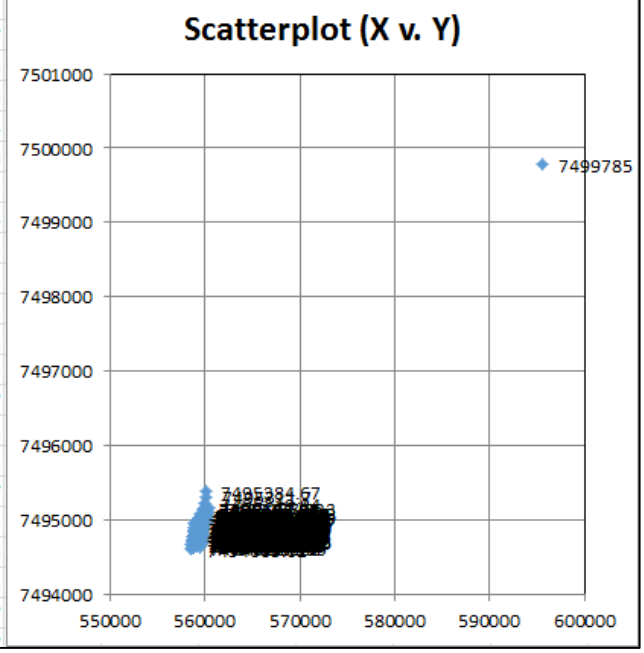
The second type of EDA covered in this guidebook is spatial EDA. As the majority of geoscience data contains a value and location associated with the sample, understanding how data behaves spatially is important. Spatial attributes can be drill collars, surface geochemical samples, or water sample collection points. When interrogating data, it is important to not only understand the data values and statistical properties of a variable but also to have confidence in the sample's location. This step provides a check using a spatially located (X, Y, and Z) dataset to quickly investigate for anomalous or erroneous locations by creating spatial data maps.

1) Arrange your data in columns with identifier, X (or easting), Y (or northing), and Z (or elevation/RL).
2) Select your X and Y data columns (columns B and C in example below).
3) Go to **Insert > Chart > Scatter** and select the chart of individual points:



4) A scatter plot of your X versus Y data should appear. Format the grid, axes, title and other attributes to your preference by **Chart Tools > Design** and **Chart Tools > Layout**.

	A	B	C	D
1	SampleID	X	Y	Z
2	AA0003	559157.3	7494791	617.307
3	AA0004	559177.9	7494656	613.528
4	AA0005	559516.4	7494833	621.489
5	AA0006	559878.6	7494890	619.054
6	AA0007	559545.4	7494629	607.586
7	AA0008	559922.2	7494712	619.713
8	AA0009	595520.4	7499785	612.4
9	BB2011	560133.7	7495385	622.03
10	BB2012	560064.9	7495312	620.3
11	BB2013	560024.6	7495220	610.71
12	BB2014	559958	7495133	621.92
13	BB2015	559919.9	7495040	616.03
14	BB10204	560437.6	7495158	607.587
15	BB10205	558803.1	7494717	625.095
16	BB10206	558795.2	7494774	622.975
17	BB10207	558788.8	7494834	630.977
18	BB10208	558774.6	7494956	658.683
19	BB10209	558777.7	7494895	644.049
20	BB10210	559037.4	7494747	620.275
21	BB10211	559024.9	7494864	628.67
22	BB10212	559021.9	7494924	634.297



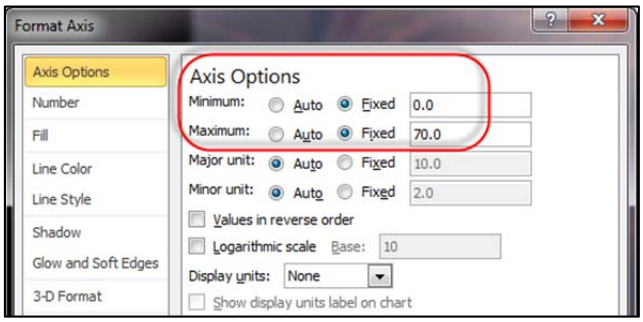
5) In the case of the data shown above, we can observe one suspicious point that is not close to the others. From a quick glance of the raw data in the table, it would be difficult to find this anomaly. Right click on the outlier point and then select **Format Data Labels**.

6) Under **Label Options**, click either the X value or Y value so you can find the point in the table. In the case above, it displays the Y-value of 7499785.
7) To find this sample, select the Y values (column C in example), then click **Find & Select > Find** to see the SampleID of this suspect Y value (SampleID = AA009). In this case, the anomalous point represents a survey station that should be deleted from the data.

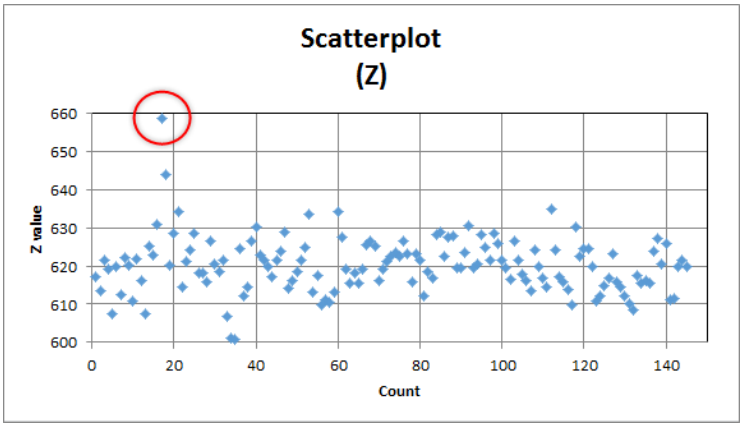
TIP: If the location of data is suspect and cannot be validated upon further investigation, it is common practice to either increase the risk rating on the data or flag it for omission during interpretation. In many cases it is better to have no data than have erroneous data biasing an interpretation. Ultimately, the geoscientist should make the call.

Next, we'll review a method of reviewing the Z or elevation values visually. This is a quick check but can save a great deal of time later on if outliers or errors are discovered. You don't want to start interpreting cross-sections with drill holes floating in space or samples located 50m in the air.

1) Select only your elevation (Z) data as shown in column D above.
2) Go to **Insert > Chart > Scatter** and select the chart of individual points as before.
3) A scatter plot of your Z versus data count will appear. Format the grid, axes, title and other features to your preference by **Chart Tools > Design** and **Chart Tools > Layout.**
4) Right click on each axis and modify the **Axis Options** so your Minimum and Maximum extents are set to **Fixed**. Then enter an appropriate value so your data is evenly spread across the graph.



5) The resulting graph will show the elevation spread for your samples. In the example below, most samples lie between ~600m and 635m but there's one, possibly two samples considerably higher (659m circled in red). This point was a drill collar entered in error that should have been 629m.

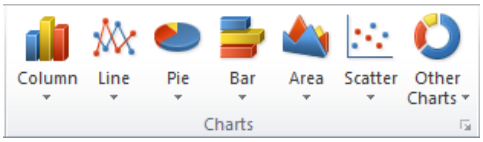


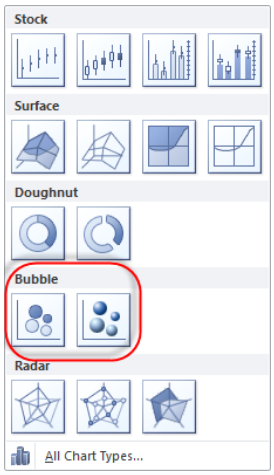
Ensuring the data is correctly located may not be clear when performing *descriptive statistics* but should be included in any EDA program. As geoscientists, we are concerned with not only the values of our data but where they are located in space and how they relate to one another or their spatial continuity. Determining errors now is a critical step prior to calculating any geostatistical parameters or understanding a variable's spatial continuity for future estimation.

Bubble Charts:

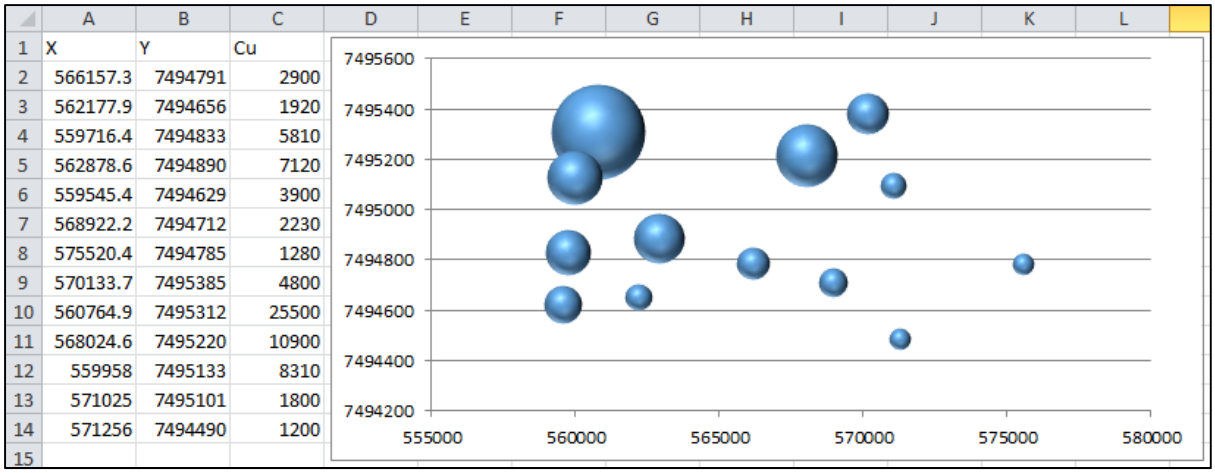
Bubble charts are effective means of representing spatial data with corresponding magnitude of data value. The features of Bubble charts are limited but useful. They are most beneficial in spatial data assessments when displaying a relatively small data population as the chart quickly becomes unreadable with too many bubbles or circles. Bubble charts require samples to have three data associated with it such as X, Y, and a numeric assay value. These are excellent in reviewing surface geochemical datasets or any other 2D data where understanding spatial trends in magnitude are required.

1) Arrange your data into three columns with X (or easting), Y (or northing), and numeric value.
2) Highlight the data of interest, then go to **Insert > Other Charts > Bubble**



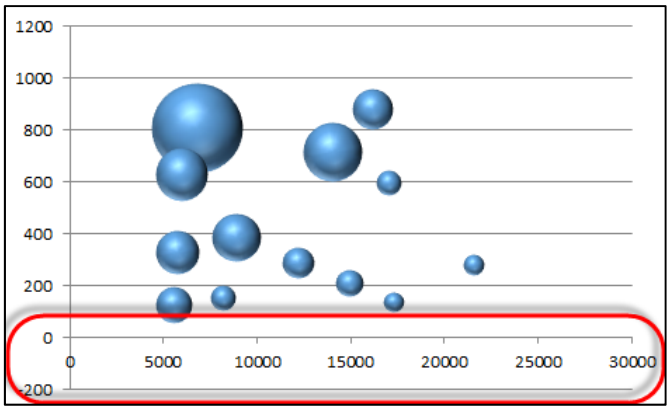


3) Select either the 2D or 3D bubbles.
4) The output chart will show the X and Y locations with a bubble of relative size for the numeric variable (in this case Cu in ppm).



	A	B	C
1	X	Y	Cu
2	566157.3	7494791	2900
3	562177.9	7494656	1920
4	559716.4	7494833	5810
5	562878.6	7494890	7120
6	559545.4	7494629	3900
7	568922.2	7494712	2230
8	575520.4	7494785	1280
9	570133.7	7495385	4800
10	560764.9	7495312	25500
11	568024.6	7495220	10900
12	559958	7495133	8310
13	571025	7495101	1800
14	571256	7494490	1200
15			

5) (Optional) The x- or y-axis may require modification if a bubble is located near either axis resulting in negative axis values.



6) Right click on an axis and select **Format Axis**.
7) Under the Axis Options, modify the Minimum to be Fixed and replace the value to 0.

The Bubble chart allows for a spatial check on data locations along with the ability to make inferences on how the sample value magnitude is related to direction and distance. Reviewing the Bubble chart above showing Cu values from a surface sampling program, the geoscientist can make the observation that Cu generally increases toward the northwest. The highest magnitude sample was collected at the end of the surface survey, therefore the Cu mineralization is open to the northwest. This information is valuable in planning a follow-up sampling campaign or providing insight into future drilling locations and directions toward increasing Cu values.

Indicator Maps:

If the size or number of data values result in a messy or difficult-to-read chart, indicator maps may be a solution. Indicators are a means of translating the data by applying a cut-off or threshold. For instance, the geoscientist may be interested in surface samples above a particular metal concentration or soil with a contaminant above a threshold. First, the data must be translated into a binary form, basically think of it as whether the individual sample value is either above or below a cut-off. Each data will have X and Y locations, then simply a "yes" or "no" as to whether it is above the cut-off of interest. So whether you chose 1 or 0,

"above" or "below", or something else, it doesn't matter as long as there are just two categories based on a cut-off. Indicator maps work best when selecting multiple cut-offs to visualize how the data behaves as cut-offs change. See below for an example using the same Cu data from above.

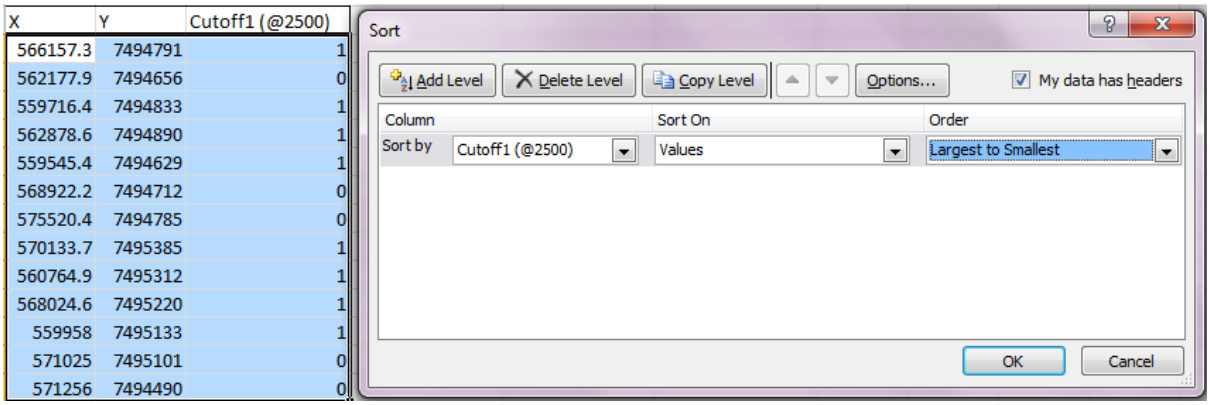
1) Arrange the data in columns with headers X (or easting), Y (or northing), and variable of interest (in this case Cu in ppm).
2) In the next column to the right, place the header for the desired cut-off or threshold values. In this case we'll consider cut-offs at 2,500 ppm, 5,000 ppm and 10,000 ppm.

	A	B	C	D	E	F
1	X	Y	Cu	Cutoff1 (@2500)	Cutoff2 (@5000)	Cutoff3 (@10000)
2	566157.3	7494791	2900			
3	562177.9	7494656	1920			
4	559716.4	7494833	5810			
5	562878.6	7494890	7120			
6	559545.4	7494629	3900			
7	568922.2	7494712	2230			
8	575520.4	7494785	1280			
9	570133.7	7495385	4800			
10	560764.9	7495312	25500			
11	568024.6	7495220	10900			
12	559958	7495133	8310			
13	571025	7495101	1800			
14	571256	7494490	1200			

3) In the first cell under the Cutoff1(@2500) cell, type " =if(C2>=2500,1,0)". This simple "if-then" statement means if the value cell C2 (Cu concentration) is equal to or larger than 2,500 ppm, then return a value of 1. If it's less than 2,500, return a 0. Drag this value down to fill in the rest of column D as necessary.
4) Repeat step 3 for Cutoff2 and Cutoff3 changing the formula appropriately to be the desired cut-off value. The resultant table should resemble:

	A	B	C	D	E	F
1	X	Y	Cu	Cutoff1 (@2500)	Cutoff2 (@5000)	Cutoff3 (@10000)
2	566157.3	7494791	2900	1	0	0
3	562177.9	7494656	1920	0	0	0
4	559716.4	7494833	5810	1	1	0
5	562878.6	7494890	7120	1	1	0
6	559545.4	7494629	3900	1	0	0
7	568922.2	7494712	2230	0	0	0
8	575520.4	7494785	1280	0	0	0
9	570133.7	7495385	4800	1	0	0
10	560764.9	7495312	25500	1	1	1
11	568024.6	7495220	10900	1	1	1
12	559958	7495133	8310	1	1	0
13	571025	7495101	1800	0	0	0
14	571256	7494490	1200	0	0	0

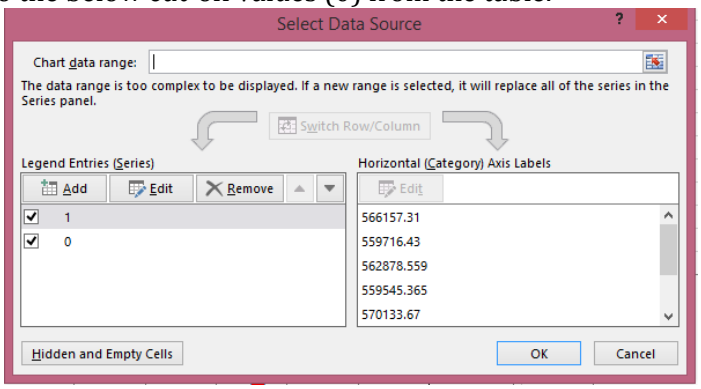
5) The next steps are a bit manual and will require copying the tables and sorting each dataset. From these, we will create three new tables each containing the X, Y, and Cut-off column of D, E, and F in separate tables.
6) Select all the data in each new table then click **Home > Sort & Filter > Custom Sort**. Sort by the Cut-off column, sort on values in order largest to smallest as shown:



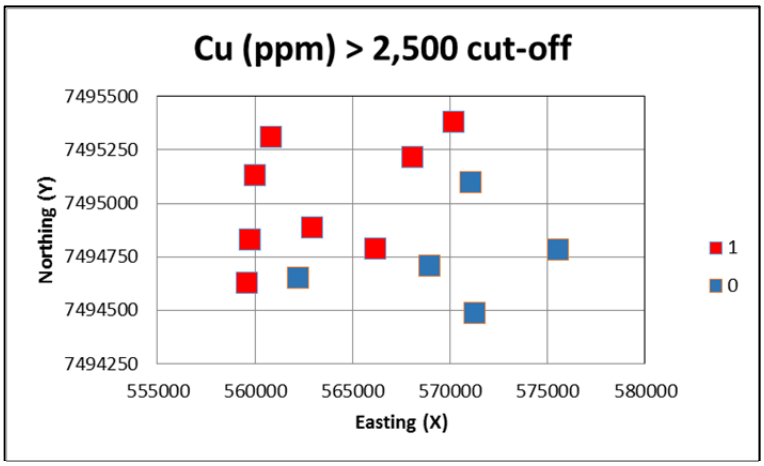
7) The resultant table should resemble the following sorted by Cutoff1:

X	Y	Cutoff1 (@2500)
566157.3	7494791	1
559716.4	7494833	1
562878.6	7494890	1
559545.4	7494629	1
570133.7	7495385	1
560764.9	7495312	1
568024.6	7495220	1
559958	7495133	1
562177.9	7494656	0
568922.2	7494712	0
575520.4	7494785	0
571025	7495101	0
571256	7494490	0

8) Next, select only the X and Y data which correspond to a Cutoff1 = 1. Then **Insert > Chart > Scatter with only Markers** plot. The resultant chart will be an X-Y scatterplot displaying the values above the cut-off.
9) Next, add the values below the cut-off. Ensure the scatterplot is selected, then go to **Design > Insert Data**. Under *Legend Entries (Series)* > **Add** with X and Y values corresponding to the below cut-off values (0) from the table.

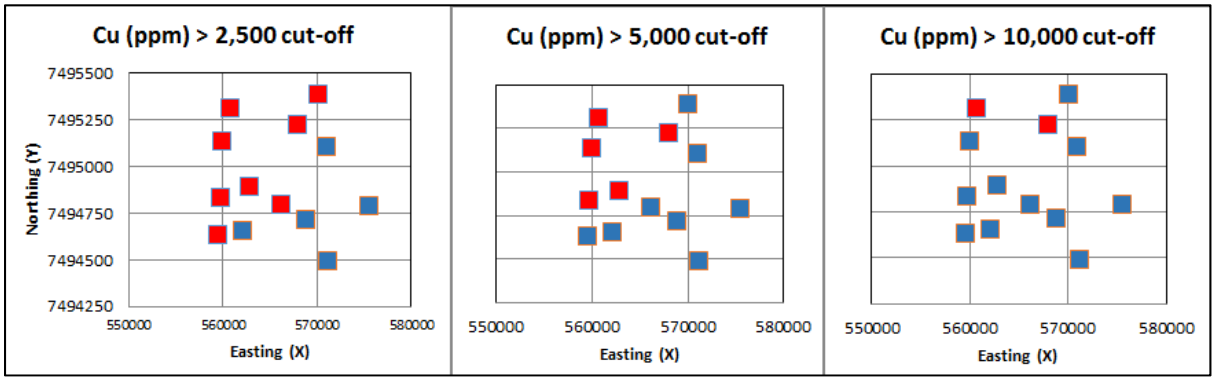


10) Modify the default marker for each data series, title, and axes. It is recommended to use a square with the above cut-off values (1) in red and the below cut-off (0) in blue. The resultant plot will resemble:



11) Repeat the previous steps to create scatterplots for Cutoff2 and Cutoff3

12) Plot all three charts together and it is easy to visualize the spatial trends in Cu concentration across the sampling area.



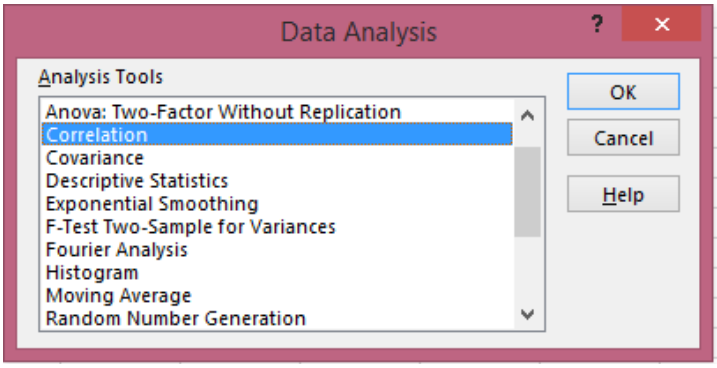
There are obvious similarities between the indicator plots and the Bubble charts though the subtle differences are valuable. In the Indicator maps, the increasing Cu trend appears to be the same general northwest direction but once the 10,000 ppm cut-off is applied, a northern trend is evident. These minor differences in visualizing data can provide insight into geological interpretations along with making future decisions on additional sampling campaigns or other evaluation activities.

10 Bivariate Analyses: Correlations and Scatterplots

The third and final type of EDA covered in this guidebook is bivariate analyses. Analyzing and understanding univariate statistics is paramount but the relationships between variables can be just as important. Bivariate statistics is simply the comparison of how two variables correlate to one another. When we discuss correlation, we're specifically talking about Pearson's correlation coefficient. This is defined as the average of the product of the z-score of the explanatory and response variables. For the sake of keeping things simple, the closer the value is to 1 the more the two variables are positively correlated (i.e. one goes up, the other goes up too). Alternatively, the closer to -1, the stronger the two variables are negatively correlated (i.e. one goes up, the other goes down).

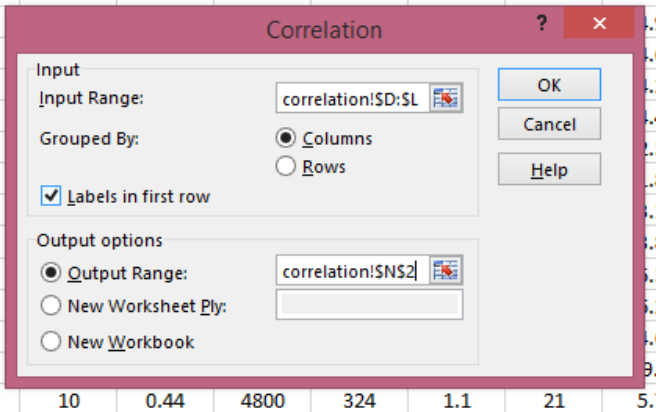
When there is a strong positive or negative correlation between two variables, this relationship can be extremely important in understanding variable behavior such as mineral alteration assemblages, enrichment/depletion of elements, or pathfinder elements for exploration. There are implications for future estimation of variables as well when there is a strong correlation between variables. The field of multivariate geostatistics is based on this correlation and ensuring it is preserved from raw data through to estimated models.

1) Go to **DATA** > **Data Analysis > Correlation** and click **OK**.



2) Select your data of interest in the Input Range. In the case presented below, there are multiple metal concentrations selected in columns D through L. These data comprise 147 samples each 1m in length. All data are presented in ppm.

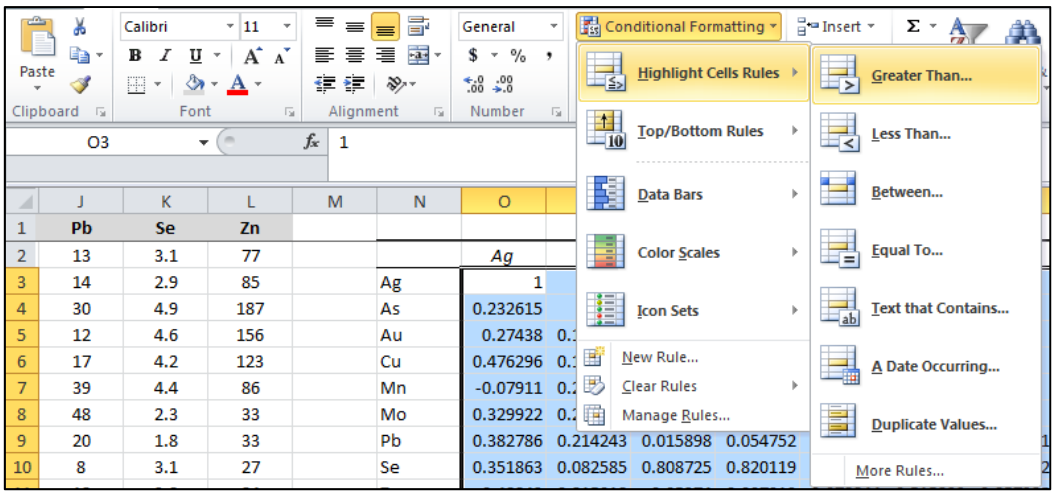
	A	B	C	D	E	F	G	H	I	J	K	L
1	HOLEID	FROM	TO	Ag	As	Au	Cu	Mn	Mo	Pb	Se	Zn
2	AA-1	44	45	1.9	13	1.53	8270	850	2.5	13	3.1	77
3	AA-1	45	46	2.1	13	1.72	9930	202	1.7	14	2.9	85
4	AA-1	46	47	2.8							.9	187
5	AA-1	47	48	3.2							.6	156
6	AA-1	48	49	3.6							.2	123
7	AA-1	49	50	3.1							.4	86
8	AA-1	50	51	2.5							.3	33
9	AA-1	52	53	1.1							.8	33
10	AA-1	54	55	1							.1	27
11	AA-1	55	56	1.3							.8	24
12	AA-1	56	57	2							.3	23
13	AA-1	57	58	2							.2	39
14	AA-1	58	59	1.2							.6	24
15	AA-1	59	60	2.6							9.1	73
16	AA-1	61	62	1.8	10	0.44	4800	324	1.1	21	5.7	31
17	AA-1	64	65	6.3	16	4.33	25500	573	2.3	41	41.3	39



3) Select the **Output options** for summary data output location.
4) The output is a table of variables across the X and Y axes with the correlation coefficient displayed in the matrix.

	Ag	*As*	*Au*	*Cu*	*Mn*	*Mo*	*Pb*	*Se*	*Zn*
Ag	1								
As	0.232615	1							
Au	0.27438	0.133862	1						
Cu	0.476296	0.125629	0.846438	1					
Mn	-0.07911	0.233094	-0.18179	-0.22677	1				
Mo	0.329922	0.241266	-0.11026	0.064711	-0.08101	1			
Pb	0.382786	0.214243	0.015898	0.054752	0.061005	0.453413	1		
Se	0.351863	0.082585	0.808725	0.820119	-0.20215	-0.03832	0.057812	1	
Zn	0.48843	0.215018	-0.05374	0.007919	0.078844	0.515233	0.857325	-0.0589	1

5) It's helpful in visualizing the data to format the output table. An easy approach is to use *Conditional Formatting*. Select the data table then go to **Home > Conditional Formatting**. There are a variety of choices but common options include **Between…** and **Greater Than…**



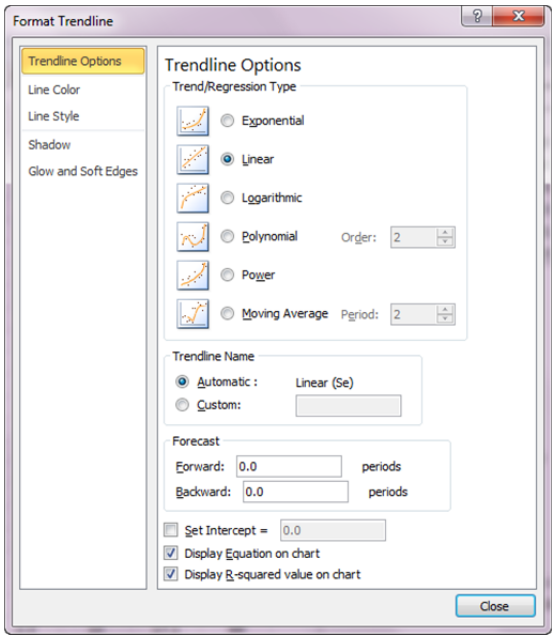
6) The example shown below uses both **Greater Than…** and **Less Than…** to assign values of > 0.6 and < -0.6 to highlight metals with strong positive or strong negative correlation coefficients.

	Ag	*As*	*Au*	*Cu*	*Mn*	*Mo*	*Pb*	*Se*	*Zn*
Ag	1.000								
As	0.233	1.000							
Au	0.274	0.134	1.000						
Cu	0.476	0.126	0.846	1.000					
Mn	-0.079	0.233	-0.182	-0.227	1.000				
Mo	0.330	0.241	-0.110	0.065	-0.081	1.000			
Pb	0.383	0.214	0.016	0.055	0.061	0.453	1.000		
Se	0.352	0.083	0.809	0.820	-0.202	-0.038	0.058	1.000	
Zn	0.488	0.215	-0.054	0.008	0.079	0.515	0.857	-0.059	1.000

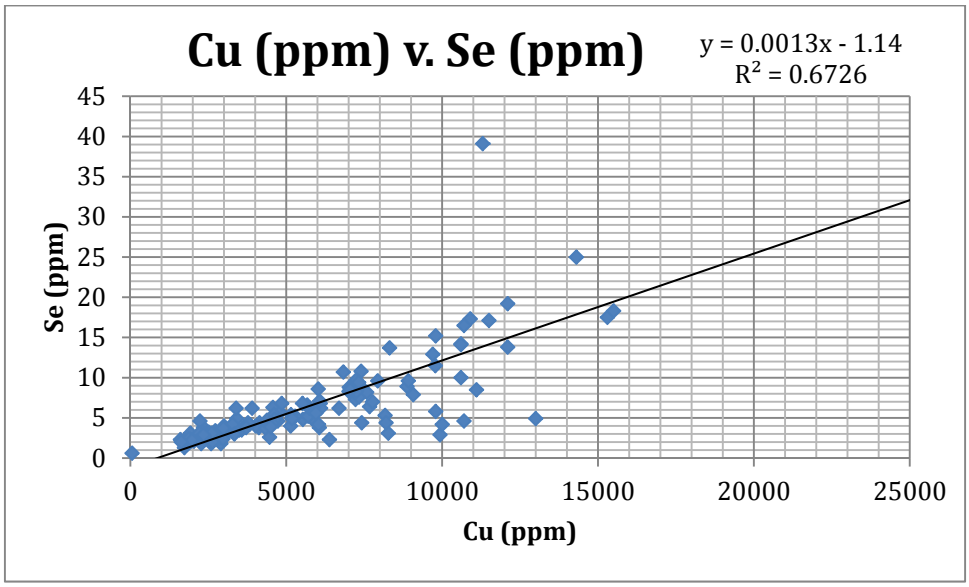
In interpreting this table, it can be stated that there are high positive correlations (>0.6) between Cu:Au, Se:Au, Cu:Se, and Zn:Pb. There are no high negative correlations (< -0.6), and there are moderate correlations of the base metals Mo:Pb, Mo:Zn, and Ag:Zn (roughly 0.4 to 0.6). Many variables are essentially not correlated to one another.

To take this analysis further, a scatter plot can be created to investigate the correlations of variables of interest. Here we'll look at Cu and Se.

1) Select the data (columns G and K in this case). Hold down the Ctrl key to select multiple columns.
2) Click **Insert** > **Charts** > **Scatter with Only Markers** chart.
3) Modify the title, axes, and legend. Alternatively, choose one of the pre-designed formats located under **Design** > **Chart Layouts**.
4) Next, we'll add a linear regression line with associated formula and R^2 value. Right click on one of the data points on the graph and select **Trendline Options**. Then select **Linear** and check the boxes to Display Equation on Chart and Display R-squared value on chart.



5) The resultant graph should resemble:



The trendline or linear regression line and R^2 value provide additional information to quantify the correlation between the two variables. R-squared the correlation coefficient (which we already calculated) squared. In the case above, that is 0.82 * 0.82 = 0.6724. Another way of looking at this is saying that about two-thirds (67.24%) of the variability of Cu can be explained by the differences in Se values or more put simply, they are strongly, but not perfect correlated in this dataset.

Now let's review the slope of the linear regression formula. In this particular case, the intercept is fairly meaningless as the Se and Cu values approach the laboratory detection limits, so few data points are available near the origin due. Instead we'll focus on the slope. In this case and as evidenced by the graph above, for every 1 ppm increase in Se value, we see a corresponding Cu increase by 1,000 ppm. Of course, the data spread or variability is high but the graph and correlation coefficient certainly confirms that there is direct relationship between Se and Cu in this dataset.

Finally, it's up to the geoscientist to understand what that actually means and the implications for your project. In the case discussed, the mineralized fluids which carried the Cu were also enriched in Se. The relationship between the two elements was a result of the same alteration event but the correlation was not a 1:1 due to the differences in country rock chemistry which preferentially concentrated more Cu in some areas versus Se. In areas with low-Cu but high-Se became interesting exploration targets to find a more favorable host rock where the high-grade Cu was potentially located.

Summary

The analysis of geological data can be a long and tedious process, but appreciating the fundamentals of EDA and performing data analysis are key steps for a geoscientist to truly understand what their data has to tell them. Unfortunately, most geoscientists are poorly exposed to, actively avoid, or simply don't appreciate the basics of EDA. Like other sub-disciplines within geology, statistics can easily become the domain of expert specialists with endless levels of sophistication and questionable practicality. However, basic data analysis skills should be as fundamental to geoscientists as petrology or structure, since today's geoscientists are commonly required to collect and manage large datasets.

Too often in industry, geoscientists attempt to jump straight into data interpretation or complicated estimation without performing the fundamental EDA. It is recommended to thoroughly understand the data prior to attempting more complicated analyses, interpretations, or estimation. Once the foundation of understanding is in place, you are encouraged to seek out additional statistical analyses along with more capable statistical software to assist with EDA.

Thank you for taking the time to read this guidebook and learn more about EDA. I hope that by following the examples and appreciating the importance of interrogating data, you have acquired some useful tools in understanding geoscience datasets. Geology is a science built on field observations, subjective hypotheses, and working theories. The industry geoscientist in the 21st century tends to be less focused on descriptive observations and more on understanding large numeric datasets. It is easier than at any time in history to collect volumes of data but seemingly more difficult to make sense of it. I hope this ten step guidebook will help geoscientists to better understand the fundamental properties of their data and thus improve their understanding of geological processes and phenomena.

Introduction to Exploratory Data Analysis (EDA) using Excel®

About Mining Geology HQ:

Mining Geology HQ was founded in 2016 to provide practical and applied information to professional geologists in an easy-to-understand manner. The company publishes online articles on a variety of mining subjects with a primary focus on mineral exploration, mine geology, and resource geology.

About the Author:

Erik C. Ronald is a geologist with nearly two decades of international industry experience across a variety of roles in mineral exploration, mine geology, and resource estimation. He holds a Bachelors of Science from the University of California - Santa Barbara, a Masters of Engineering from the Colorado School of Mines, and a Graduate Certificate in Geostatistics from Edith Cowan University. Mr. Ronald is a registered professional geologist in the U.S. State of Wyoming.

Front Cover:

The illustration is a Box and Whisker plot and frequency distributions displaying the associated statistical properties of each chart.