Morgan Stanley

Data Quality

05/23/2023

David Daniel

david.daniel@morganstanley.com

Why is Data Quality Important?

All data is not created equal, some elements are more critical than others.

Once data is created, we have no control over how it is used, the best we can do is ensure it is as good as possible.
Elements that are critical should be identified, documented, and tested against business logic along critical data quality dimensions[1].

Completeness

All the necessary elements contain values.

Uniqueness

Nothing will be recorded more than once based upon how that thing is identified.

Timeliness

The degree to which data represent reality from the required point in time.

Validity

Data are valid if it conforms to the syntax (format, type, range) of its definition.

Accuracy

The degree to which data correctly describes the "real world" object or event being described.

Consistency

The absence of difference, when comparing two or more representations of a thing against a definition.

Items deemed critical, should have a measurable validity score[2].

[1]: DAMA UK Working Group on "Data Quality Dimensions" (2013).
[2]: Bhimani & Daniel 2016

Data Quality: Completeness Example

Background

Let's say we classify our locations in different ways:

- Managed – Locations that are leased or owned
- Unmanaged – Locations that are owned or managed by out outsource providers

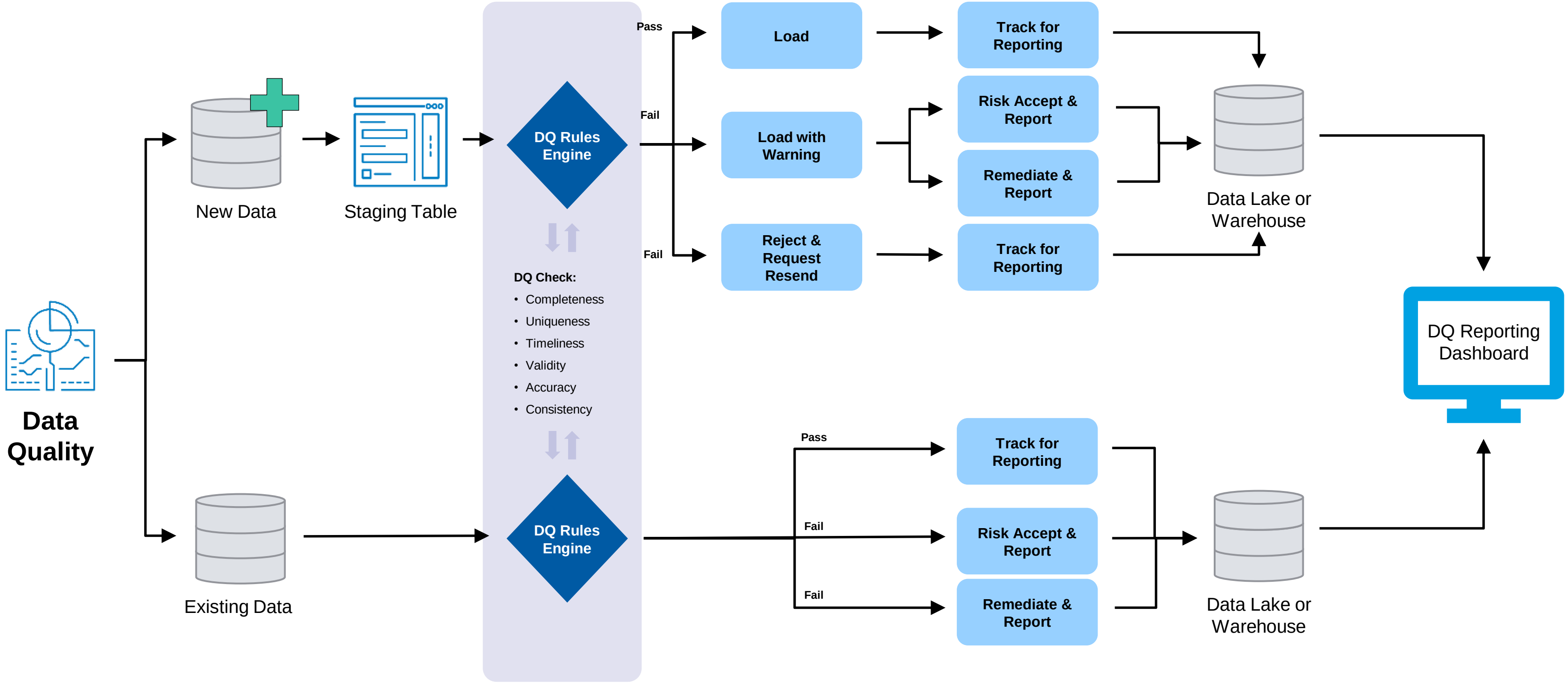
Challenge

- This data is stored in the source system in a user defined attribute (Location Type), but it is not a required field.
- Potentially, Location Type is sent to our warehouse with no data.
- If a user report is requested with Location Type as a required field, the report will display information by:
  - Managed
  - Unmanaged
  - NULL

Solution

- Remediate data in both systems (temporary solution)
- Designate Location Type as a Critical Data Element in your DQ environment
- Establish a completeness check to insure it is always present.

DQ Rule and Reporting Strategy for Critical Data Elements



DQ Check:

- Completeness
- Uniqueness
- Timeliness
- Validity
- Accuracy
- Consistency

Conscious Effort to Not Overstate Data Quality[1]

Calculating Data Quality: Items deemed critical, should have a measurable validity score.

Principles

- Fundamentally, the heart of the DQ calculation is:

$$\frac{Number\ of\ Successful\ Records}{Total\ Number\ of\ Records}$$

- Only count the records where rules apply
  - Rather than strictly returning Pass or Fail, the DQ rule should return NA, where the rule does not apply.

Example

- Assume a data file of 10,000 records
- 5000 records where *Interest Rate Type = FIXED*
- 5000 records where *Interest Rate Type = VARIABLE*
- There is a DQ Rule: *IF Interest Rate Type = FIXED THEN Interest Rate Spread != 0 – This fails 600 times*

Counting ALL records

Without NA consideration:

$$\frac{9{,}400}{10{,}000} = 94\%$$

Counting only FIXED records

With NA consideration:

$$\frac{4{,}400}{5{,}000} = 88\%$$

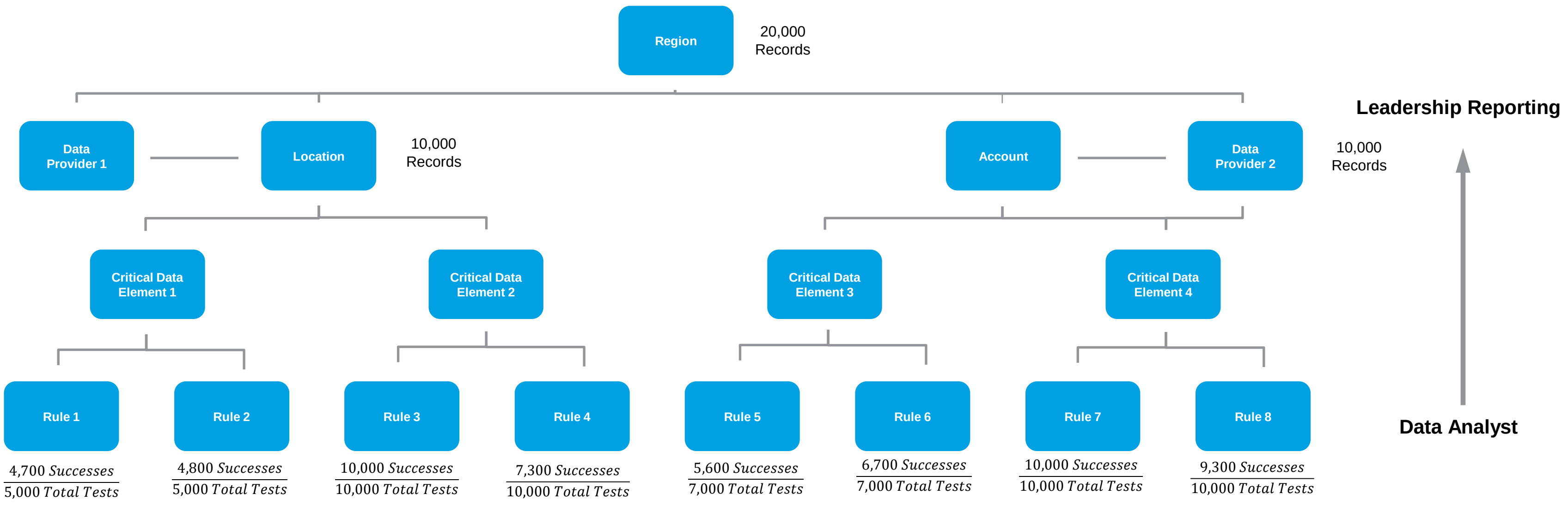
The DQ percentage changes but, there were always 600 errors.

Without the NA, the denominator*, Total Number of Records*, can grow to be very large while the number of *Failed Records* remains constant.

[1]: Bhimani & Daniel 2016

Aggregation and Reporting*

Different levels of the organization require different information about data quality



*All counts in this presentation are fabricated for demonstration purposes

Aggregation and Reporting (cont.)

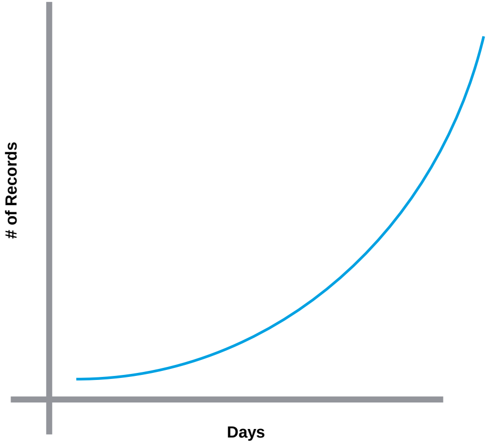
Example based on previous slide:

- Assume 1 file from each data provider every weekday and each file has:
  - 10,000 records
  - 2 critical data elements
  - 4 DQ Rules

- File 1 (Location)
  - 5,000 + 5,000 + 10,000 +10,000 = 30,000 tests per day

- File 2 (Account)
  - 7,000 + 7,000 + 10,000 +10,000 = 34,000 tests per day

- 5 files from each provider per week



30,000 x 5 = **150,000**
and
34,000 x 5 = **170,000**

Totaling = **320,000**
results per week
and
16,640,000 Results per year

- As seen, the amount of data grows drastically over time; this could place performance challenges on reporting.

- As such, planning your desired reporting in advance is critical. It is helpful to use database processes to pre-aggregate the data for dashboards.

Data Quality Reporting Basics

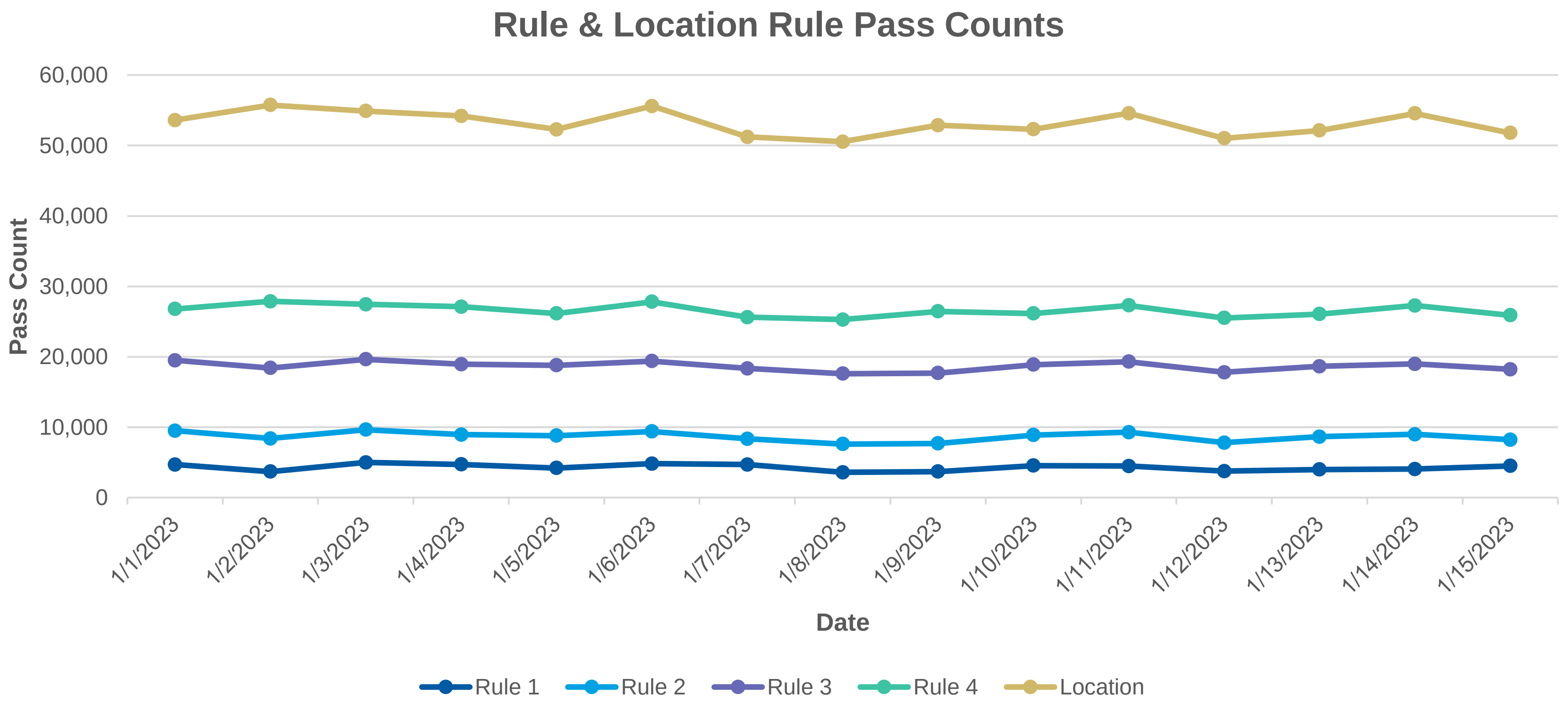
Aggregation Level	Calculation	Score
Rule 1	4,700 / 5,000	**94.0%**
Rule 2	4,800 / 5,000	**96.0%**
Rule 3	10,000 / 10,000	**100.0%**
Rule 4	7,300 / 10,000	**73.0%**
Rule 5	5,600 / 7,000	**80.0%**
Rule 6	6,700 / 7,000	**95.7%**
Rule 7	10,000 / 10,000	**100.0%**
Rule 8	9,300 / 10,000	**93.0%**
Critical Data Element 1	(4,700 + 4,800) / (5,000 + 5,000)	**95.0%**
Critical Data Element 2	(10,000 + 7,300) / (10,000 + 10,000)	**86.5%**
Critical Data Element 3	(5,600 + 6,700) / (7,000 + 7,000)	**87.9%**
Critical Data Element 4	(10,000 + 9,300) / (10,000 + 10,000)	**96.5%**
Location	(4,700 + 4,800 + 10,000 + 7,300) / (5,000 + 5,000 + 10,000 + 10,000)	**89.3%**
Account	(5,600 + 6,700 + 10,000 + 9,300) / (7,000 + 7,000 + 10,000 + 10,000)	**92.9%**
Region	(4,700 + 4,800 + 10,000 + 7,300 + 5,600 + 6,700 + 10,000 + 9,300) / (5,000 + 5,000 + 10,000 + 10,000 + 7,000 + 7,000 + 10,000 + 10,000)	**91.3%**

Data Quality Reporting Basics (cont.)

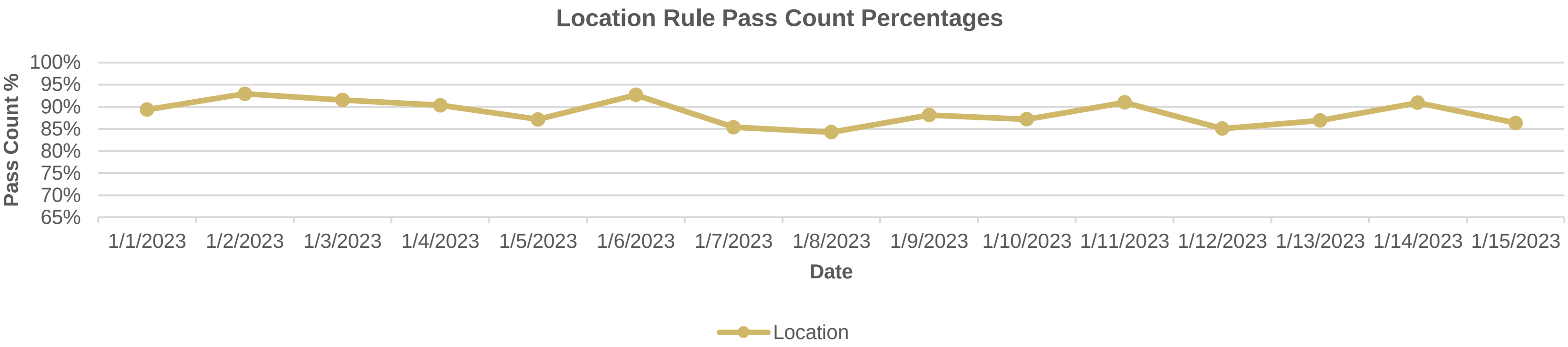
While a score is useful, it is also only a snapshot in time. It may also be useful to look at trending and volatility measures.

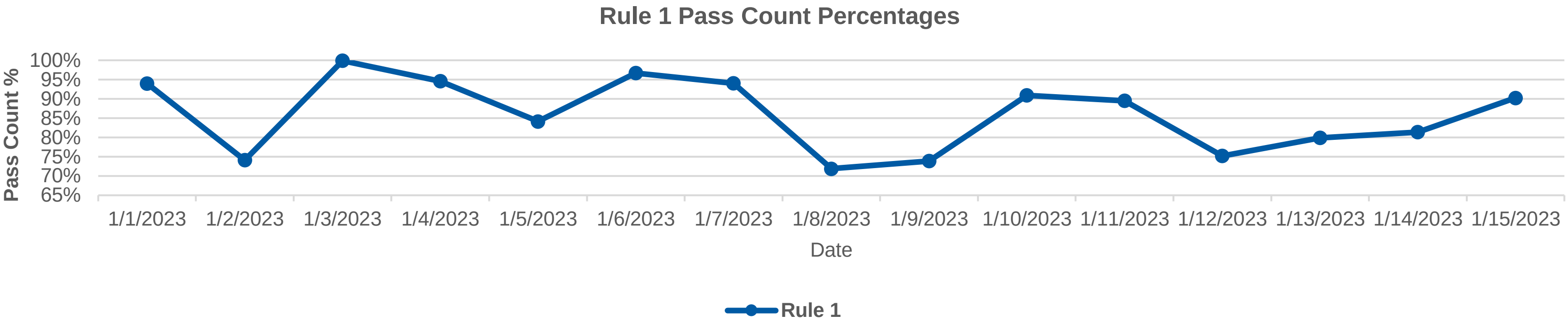
	1/1/2023	1/2/2023		1/14/2023	1/15/2023	10 Day Moving Average	10 Day Moving Average %	Monthly Avg To Date	Monthly Avg To Date %
Rule 1	4,700	3,706		4,070	4,512	4,217.5	84.4%	4,301.5	86.0%
Rule 2	4,800	4,698		4,925	3,715	4,299.5	85.4%	4,377.7	87.6%
Rule 3	10,000	10,000		10,000	10,000	10,000.0	100.0%	10,000.0	100.0%
Rule 4	7,300	9,471		8,281	7,671	7,799.7	78.5%	7,901.8	79.0%
Location	26,800	27,875	• • •	27,276	25,898	26,316.6	87.8%	26,580.9	88.6%

Data Quality Reporting – Trending



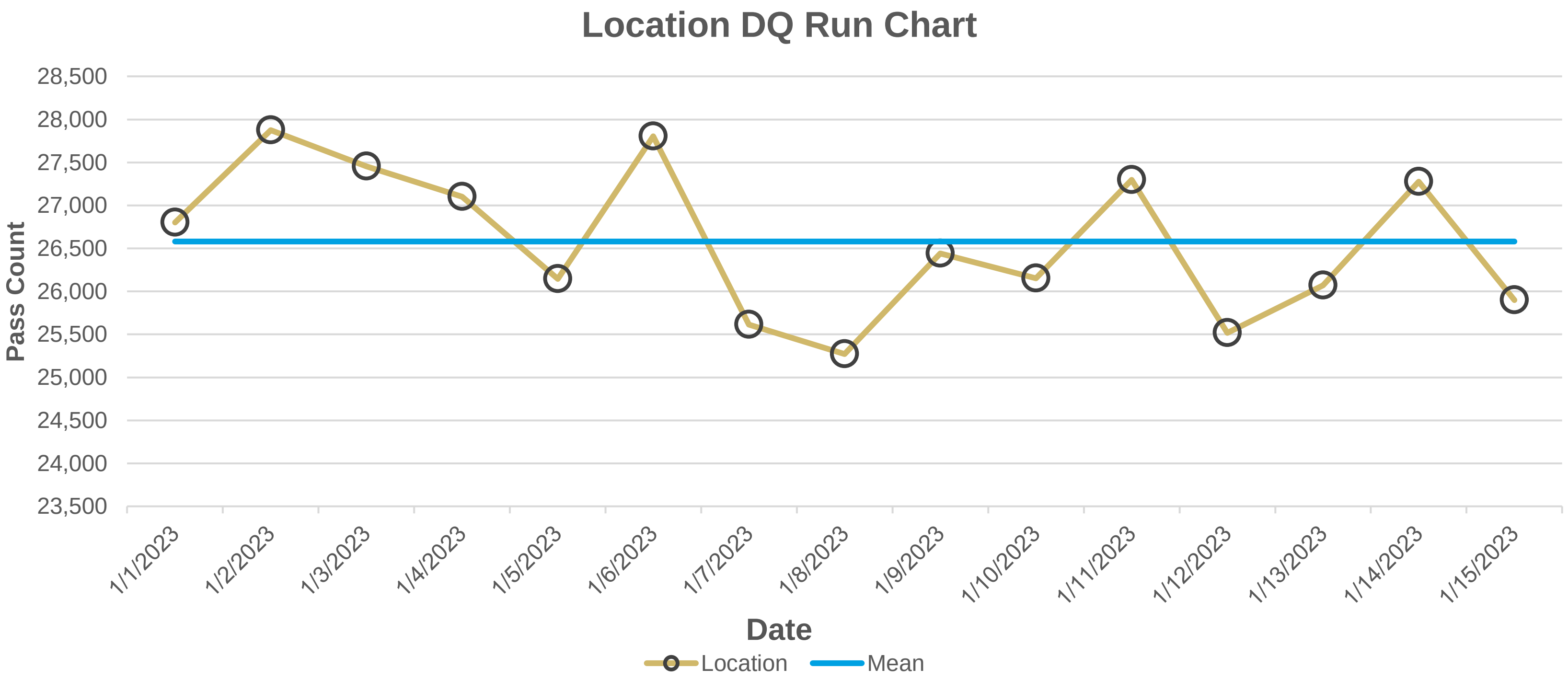
Data Quality Reporting – Trending (cont.)





Data Quality Run Chart[1]

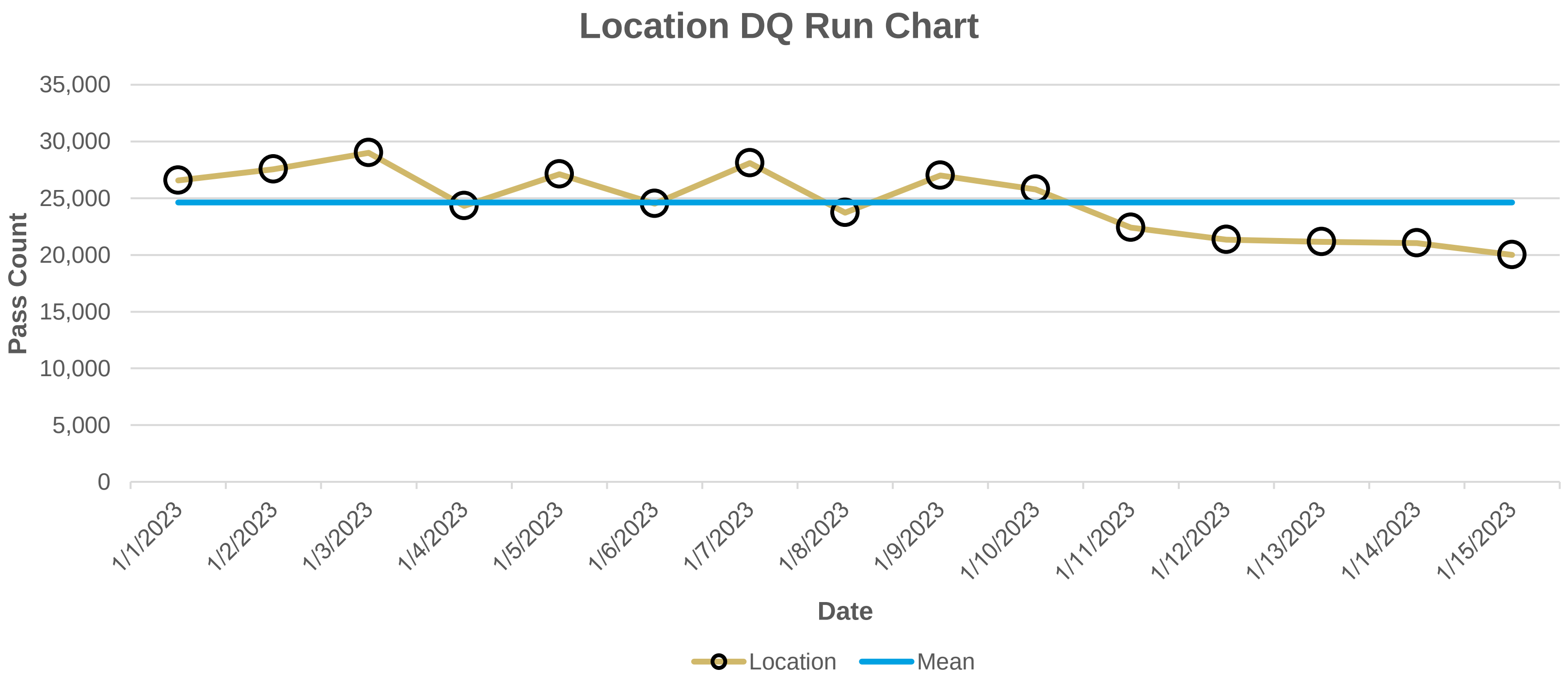
By all appearances, this is a stable process; all the observations are randomly distributed about the mean.



[1]: Christopher J. Wild & George A. F. Seber (1999)

Data Quality Run Chart (cont.)

This is an unstable process; the observations are drifting downwards.



Data Quality Control Chart

The *x*-chart is used to look for changes in the average value of *X*-measurements as time goes on. As the measured characteristic of the process may not be Normally distributed, we make use of the Central Limit effect by working with sample means instead of individual *X*-values so that we are working with quantities that have a distribution that is closer to Normal. Typically, data are collected in *at least* 20 subgroups of size 3 to 6 (typically 5) measurements and the mean of each of subgroup is computed[1].

						Subgroup		
Dates	**Day 1**	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**Mean**	**Range**	**Std Dev**
1/2 - 1/6	25,103.0	27,875.0	27,455.0	27,101.0	26,144.0	26,736.0	2,772.0	1,114.1
1/9 - 1/13	24,654.0	20,643.0	20,985.0	25,517.0	21,873.0	22,734.0	4,874.0	2,213.8
1/16 - 1/20	28,863.0	26,962.0	24,212.0	28,606.0	20,167.0	25,762.0	8,696.0	3,634.9
1/23 - 1/23	20,971.0	27,184.0	26,991.0	29,603.0	28,250.0	26,600.0	8,632.0	3,313.6
1/30 - 2/3	26,461.0	28,285.0	23,461.0	23,502.0	22,595.0	24,861.0	5,690.0	2,409.6
2/6 - 2/10	21,583.0	27,815.0	23,157.0	28,757.0	22,126.0	24,688.0	7,174.0	3,349.8
2/13 - 2/17	28,744.0	27,248.0	20,544.0	21,003.0	29,126.0	25,333.0	8,582.0	4,224.1
2/20 - 2/24	26,589.0	24,937.0	22,255.0	20,111.0	25,660.0	23,910.0	6,478.0	2,667.7
2/27 - 3/3	29,727.0	23,412.0	25,177.0	24,473.0	20,706.0	24,699.0	9,021.0	3,284.8
3/6 - 3/10	22,196.0	28,359.0	29,497.0	29,568.0	23,739.0	26,672.0	7,372.0	3,458.7
3/13 - 3/17	21,134.0	21,131.0	25,987.0	27,104.0	29,830.0	25,037.0	8,699.0	3,828.8
3/20 - 3/24	21,117.0	27,230.0	26,567.0	24,454.0	26,064.0	25,086.0	6,113.0	2,444.3
3/27 - 3/31	21,529.0	20,326.0	24,586.0	28,273.0	25,525.0	24,048.0	7,947.0	3,182.8
4/3 - 4/7	23,318.0	22,246.0	23,320.0	22,822.0	20,017.0	22,345.0	3,303.0	1,374.3
4/10 - 4/14	25,702.0	24,298.0	25,610.0	21,815.0	23,089.0	24,103.0	3,887.0	1,668.0
4/17 - 4/21	26,117.0	24,152.0	28,944.0	26,389.0	20,984.0	25,317.0	7,960.0	2,961.3
4/24 - 4/28	23,685.0	24,060.0	29,006.0	24,386.0	26,577.0	25,543.0	5,321.0	2,239.0
5/1 - 5/5	26,650.0	24,298.0	27,217.0	23,832.0	28,259.0	26,051.0	4,427.0	1,909.9
5/8 - 5/12	20,055.0	29,099.0	28,903.0	23,431.0	25,452.0	25,388.0	9,044.0	3,821.0
5/15 - 5/19	26,641.0	27,846.0	28,688.0	25,193.0	21,660.0	26,006.0	7,028.0	2,762.7
					Column Mean =	24,995.0	6,631.0	2,795.0
						x double bar	**r-bar**	**s-bar**

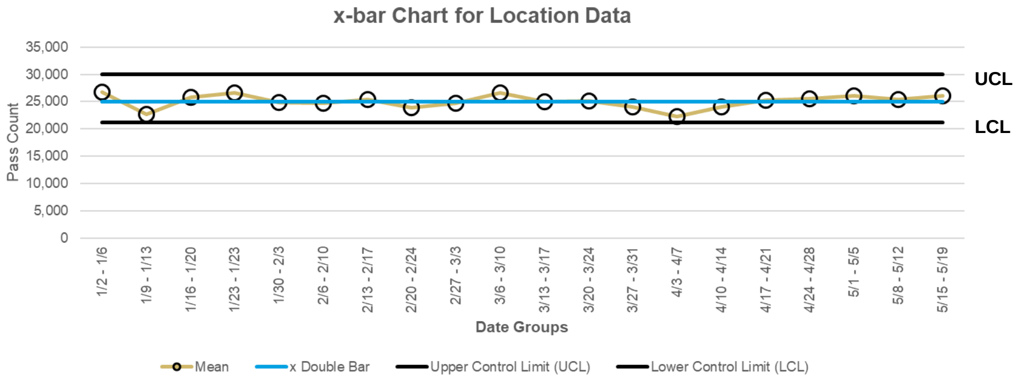
[1] Christopher J. Wild & George A. F. Seber (1999)

Data Quality Control Chart (cont.)

Because the sample means (x-Bar) are normally distributed, we can say that 99.7% of the observed DQ means should fall within 3 standard deviations of the mean. So, anything outside -3 standard deviations from the mean is a problem. Though, in practice, you may decide that -3 standard deviations is too wide and shrink the boundary to meet your business need.

In the x-Bar chart[1] the **subgroup means** are plotted in relation to 3 other lines:

- The Center Line, x Double Bar – This is the mean of all the sample means
- Upper Control Limit (UCL)– This the set at a perfect DQ score (no defects)
- Lower Control Limit (LCL)– This is -3 Standard deviations from x Double Bar



The chart above shows that the average level of the Location Data is stable; the data is withing the control limits.

$UCL = 30{,}000$ – No DQ defects

$$Center\ Line = x\ Double\ Bar$$

$$LCL = x\ Double\ Bar\ - 3\,\overset{\wedge}{\sigma}_{x-bar} = 21170$$

- $\overset{\wedge}{\sigma}_{x-bar} = \frac{1}{d_2}\frac{r-bar}{\sqrt{n}} = \frac{1}{2.3259}\frac{6631}{\sqrt{5}} = 1275$ [1]

Note: d_2 is a constant that can be found on the internet. For a normal distribution, $\frac{r-bar}{d_2}$ is an unbiased estimate of σ, the standard deviation of the actual population.

[1]: Christopher J. Wild & George A. F. Seber (1999)

Data Quality Control Chart (cont.)

In using the R-chart, subgroup ranges lying outside the control limits indicate that the process is “out of control”.

In the R-chart the **subgroup ranges** are plotted in relation to 3 other lines:

- The Center Line, r–bar – This is the mean of all the subgroup ranges
- Upper Control Limit (UCL) – This is +3 Standard deviations from r-bar and is calculated as follows[1]
- Lower Control Limit (LCL) – This is -3 Standard deviations from r-bar and is calculated as follows[2]:

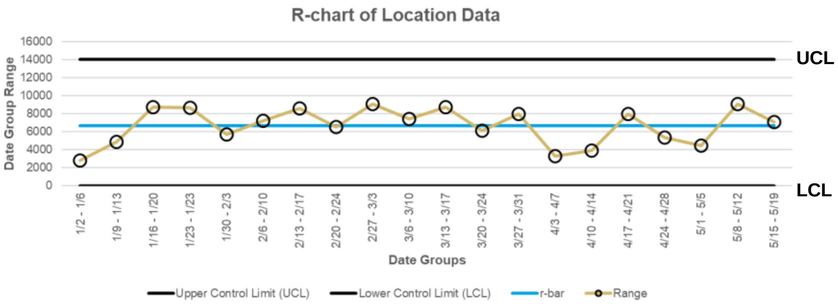
$$UCL = \mathrm{r} - bar + 3\,\overset{\wedge}{\sigma}_{r-bar}\quad D_4 * \mathrm{r} - bar = 2.1145 * 6631 = 14{,}021.58336 \cong 14{,}022$$

$$Center\ Line = \mathrm{r} - bar$$

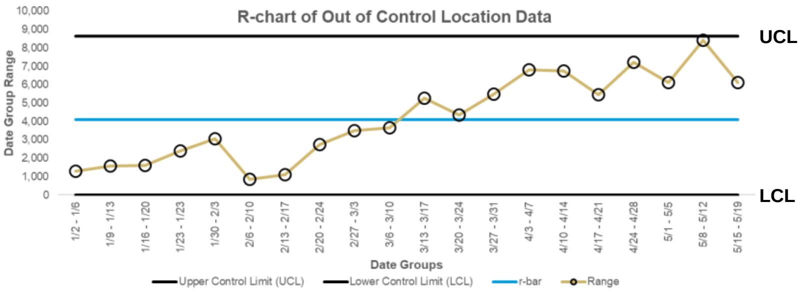
$$LCL = \mathrm{r} - bar - 3\,\overset{\wedge}{\sigma}_{r-bar}\quad D_3 * \mathrm{r} - bar = 0\ * 6631 = 0$$

Note: D_3 & D_4 are bias correction constants and that can be found on the internet.

An x-Bar chart focuses attention on the constancy of average level (μ) and is not good at detecting changes in variability (σ). An R-chart (or range chart) is specifically designed for detecting changes in variability.[1]



The R-chart below indicates that the process below is out of control. Notice how the range is increasing over time.[1]



[1] Christopher J. Wild & George A. F. Seber (1999)

[2] Ibid.

Bibliography

Adobe Acrobat Document

DAMA UK Working Group on “Data Quality Dimensions”.(2013). *THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT Defining Data Quality Dimensions* [White Paper]

Adobe Acrobat Document

Umesh Bhimani, & David Daniel (2016). *Data Quality: Solutions and Pitfalls* [White Paper]

Adobe Acrobat Document

Christopher J. Wild & George A. F. Seber (1999). Chance Encounters: A First Course in Data Analysis and Inference 1st Edition. Chapter 13 [Textbook]

Special thanks to Gabrielle Cha of Morgan Stanley for help with the methodology and contriving the data for the Control Charts, and Adem Islami of JLL for formatting and presentation help.

Contact Info:
David Daniel
david.daniel@morganstanley.com