Simple Random vs. Cluster

By Arnold Miller, Sr. SDET (SW QA-Tester)

ASQ Boulder (Section 1313)

Boulder, CO, USA

23 Jan 2020

Sampling Methods

Sampling method refers to the way that observations are selected from a **population** to be in the **sample** for a **survey sample**

https://stattrek.com/survey-research/sampling-methods.aspx

The reason for conducting a sample survey is to **estimate** the **value** of some **attribute** of a **population**.

- **Population parameter**. A population parameter is the true value of a population attribute.
- **Sample statistic**. A sample statistic is an estimate, based on sample data, of a population parameter

Non-Probability Sample Methods

Do not know the probability that each population element will be chosen,

Cannot be certain that each population element has a non-zero chance of being chosen.

Two Main types are:

- **Voluntary sample**. Made up of people who self-select into the survey. Often, these folks have a strong interest in the main topic of the survey.
- **Convenience sample**. Made up of people who are easy to reach.

Probability Sample Methods

Each population element has a known (non-zero) chance of being chosen for the sample.

Main Sampling Methods

- **Simple Random**
- **Stratified**
- **Cluster**
- **Multistage**
- **Systematic Random**

Simple Random Sampling

Property

- The population consists of N objects.
- The sample consists of n objects.
- All possible samples of n objects are equally likely to occur

Example: Lottery method.

- Each of the N population members is assigned a unique number.
- The numbers are placed in a bowl and thoroughly mixed.
- A blind-folded researcher selects n numbers.
- Survey only members that have the selected numbers

Stratified Sampling

Property

- Population is divided into groups (strata), based on some characteristic.
- Then, within each group, a probability sample (like: Simple Random Sampling) is selected.
- In stratified sampling, the groups are called **strata**.
- The sample includes **elements** from **each stratum**.

Example: US National Elections

- Divide the population into groups or strata, based on geography like US States
- Within each stratum use simple randomly select survey respondents.

Cluster Sampling

Property

- Every member of the population is assigned to one, and only one, group.
- Each group is called a **cluster**.
- A sample of clusters is chosen, using a probability method (like: Simple Random Sampling).
- Survey only individuals within sampled clusters
- the sample includes **elements** only from **sampled clusters**

Example: Items in Group

- Each Group has 10 times.
- Select Groups via Simple random sampling
- Survey all or simple random items in Selected Group

Multistage Sampling

Property

- Select a sample by using combinations of different sampling methods.

Example

- Stage 1, Use cluster sampling to choose clusters from a population.
- Stage 2, Use simple random sampling to select a subset of elements from each chosen cluster for the final sample.

Systematic Random Sampling

Property

- Create a list of every member of the population.
- From the list, we randomly select the first sample element from the first k elements on the population list.
- Thereafter, we select every kth element on the list.
- Every possible sample of n elements is not equally likely (not a simple random sample)

Example

- Have 24,000 items in some order
- Random select first item from first 50 items
- Second select item is first random plus 50; Third select item is first random plus 100; etc.
- Survey has total 480 items (Which is 24,000 / 50)

Confidence Interval

Best for Simple Random Samples

Plus or Minus error value reported based on

Confidence level: Percentage certain for the interval. Normal 95% or 99%

Sample Size: numbers of items surveyed. The more the better but no linear

Population Size: Total number that could be surveyed

Percentage: Likely outcome via population size

https://www.surveysystem.com/sscalc.htm

Sample	Population	Percent	95% Cnf	99% Cnf
600	6,300	50%	3.81%	5.01%
600	63,000	50%	3.98%	5.24%
600	630,000	50%	4.00%	5.26%
600	6,300,000	50%	4.00%	5.27%

Example: Compare Probability Sample Methods

On-Demand TV Shows (about 6,300 total population)

Web Site displays TV Shows in sort groups of 100 (like: Popular, **Title**, Critics, Date added, Relevance)

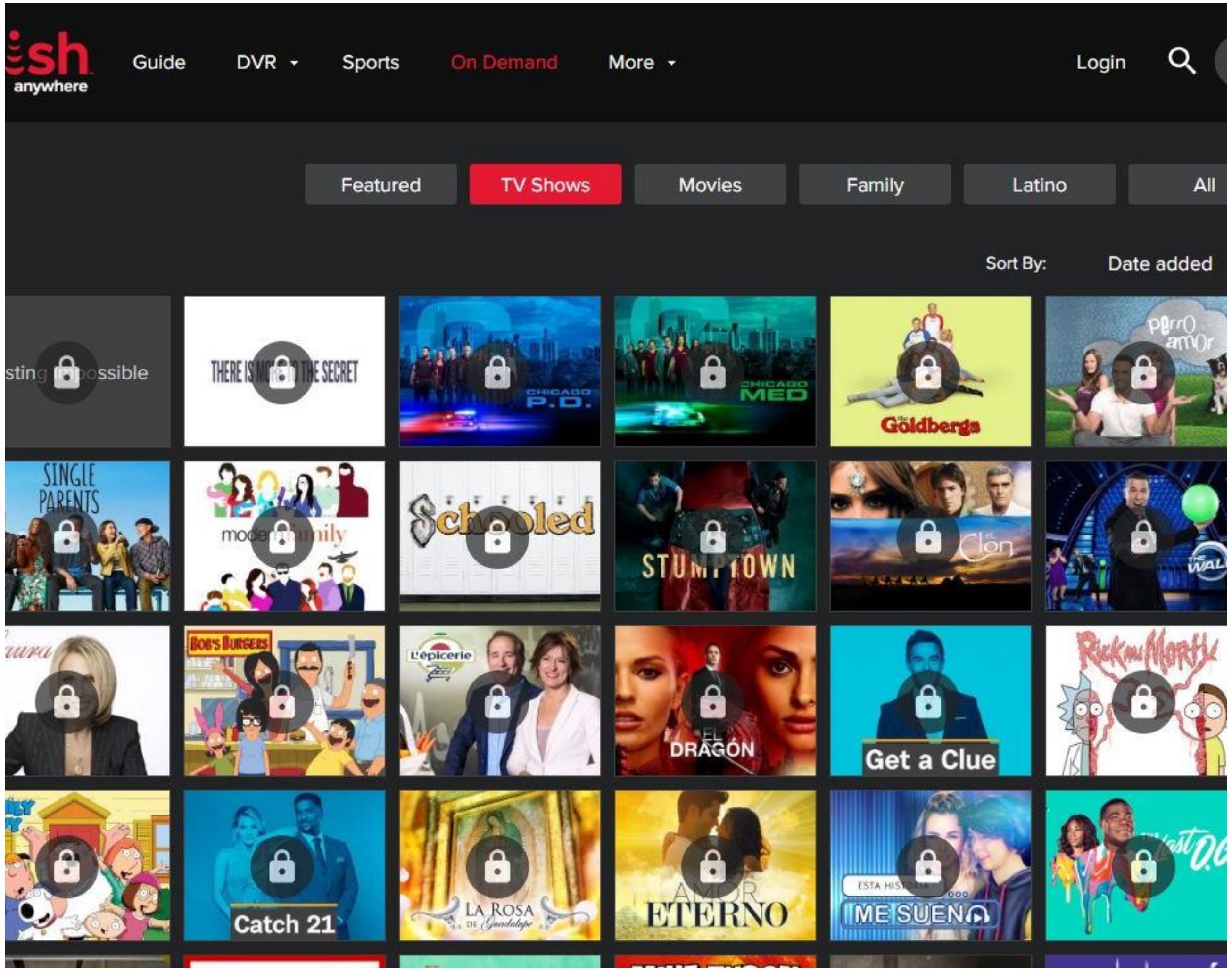
Other Groupings (Genre, View Rating, Start Letter)

Evaluate TV Shows for attributes: Description, View Rating, Genre, Network, Image

Percentage has all these attributes

Percentage missing each one attribute

- No Description
- No View Rating
- No Genre
- No Network
- No Image



Example: Sampling Models

Total Population (sort: Title) 63 groups of 100, just like Web site

Simple Random 600 individual items: Execution Baseline

Stratified Start Letter with pro-rated Simple Random individual items

- 600 items pro-rated via Strata population with minimum 1 per Strata

Cluster 60 Simple Random groups of 10. Evaluate all items in group

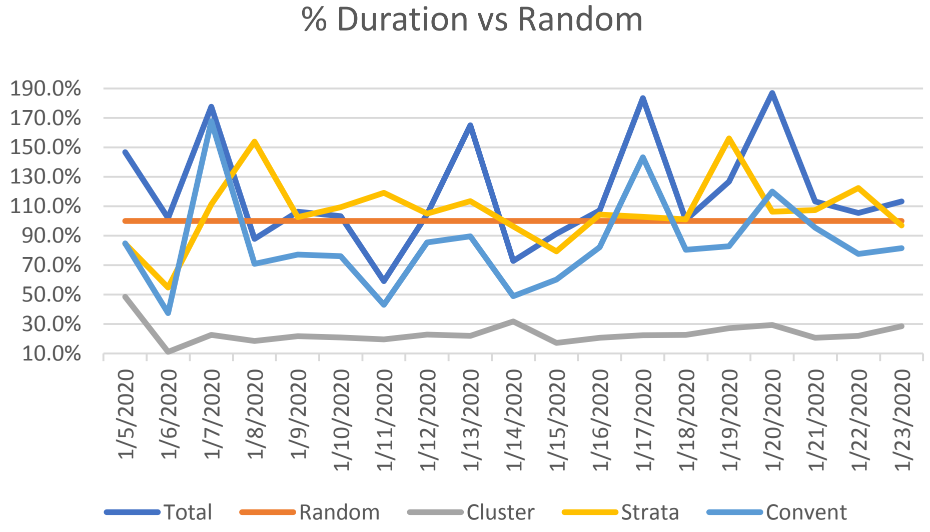
Convent first 100 each Sort; first 50 each Genre, View Rating; first 25 each Start Letter

- Genre (Action, Comedy, Drama, How-to, Family, Sci-Fi, Sports, etc.)
- View Rating (G, PG, PG-13, R, NR, TV-Y, TV-G, TV-PG, TV-14, TV-MA)

Execution time vs Random Base line

Sampling Methods (Summary)

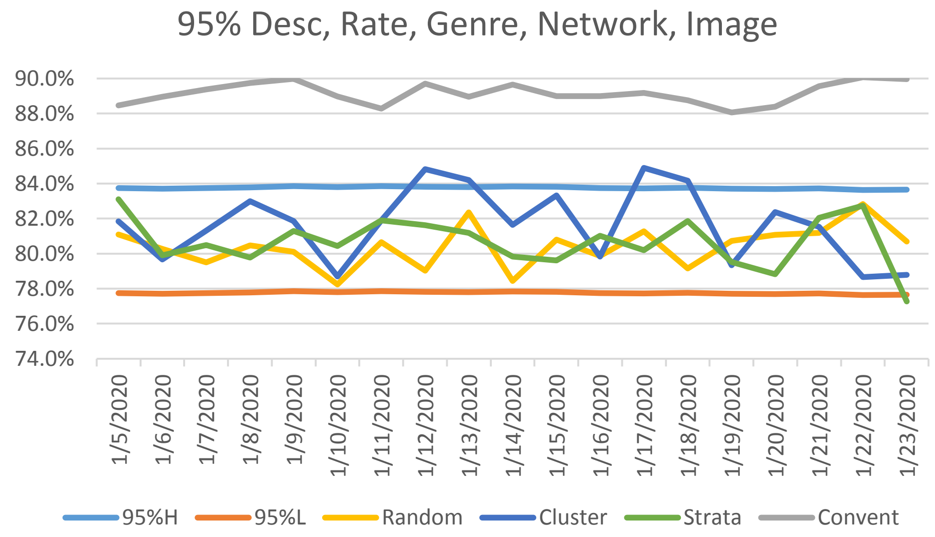
- **Total Populate** (Sort: Title) All 6300 items (120% baseline)
- **Simple Random** (Sort: Title) Execution Baseline (average 5 min 30 sec)
- **Stratified** (Sort: Title) 600 random pro-rated Start Letter (110% baseline)
- **Cluster** (Sort: Title) 60 Simple Random groups of 10; (25% baseline)
- **Convent** 100 each Sort; 50 each Genre, View Rating; 25 each Letter (85% baseline)

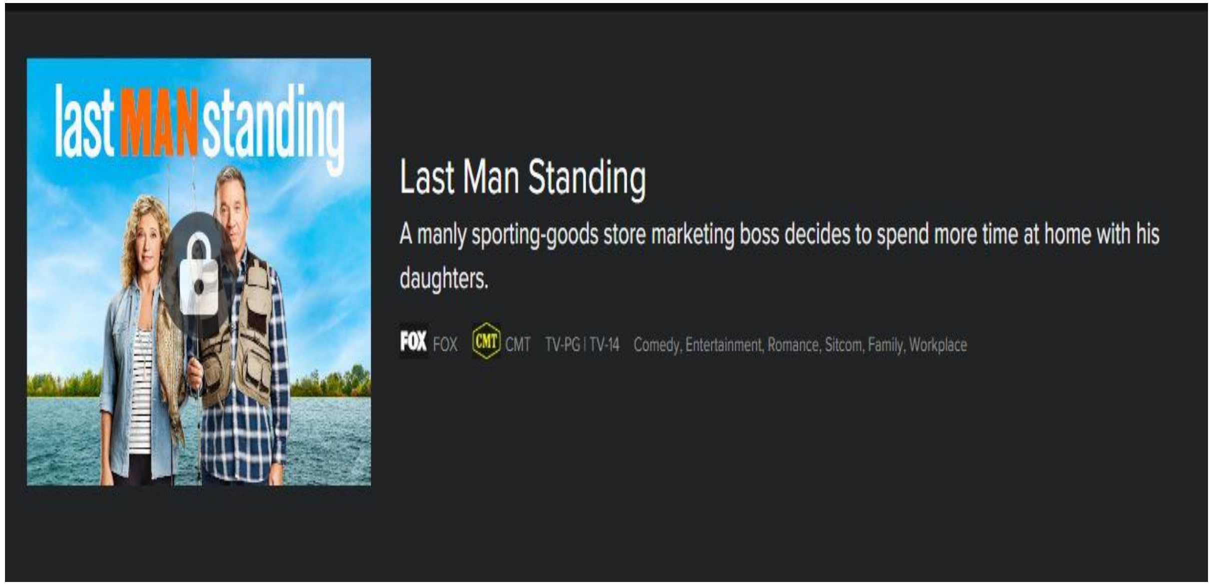


Example: Attribute – Have All

Historically: 80.8% have Description, View Rating, Genre, Network, Image

Sample	Population	Percent	95% Cnf	99% Cnf
600	6,300	80.8%	3.00%	3.95%





Last Man Standing

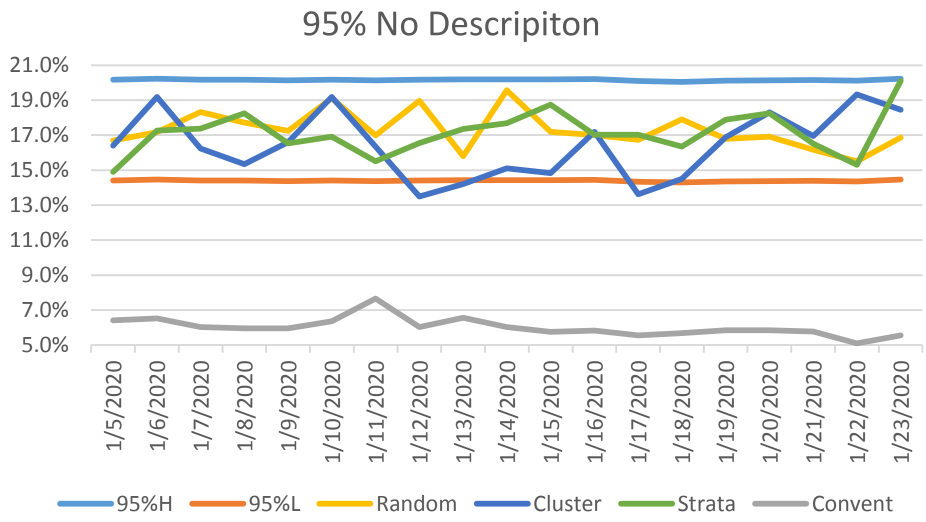
A manly sporting-goods store marketing boss decides to spend more time at home with his daughters.

FOX FOX CMT CMT TV-PG | TV-14 Comedy, Entertainment, Romance, Sitcom, Family, Workplace

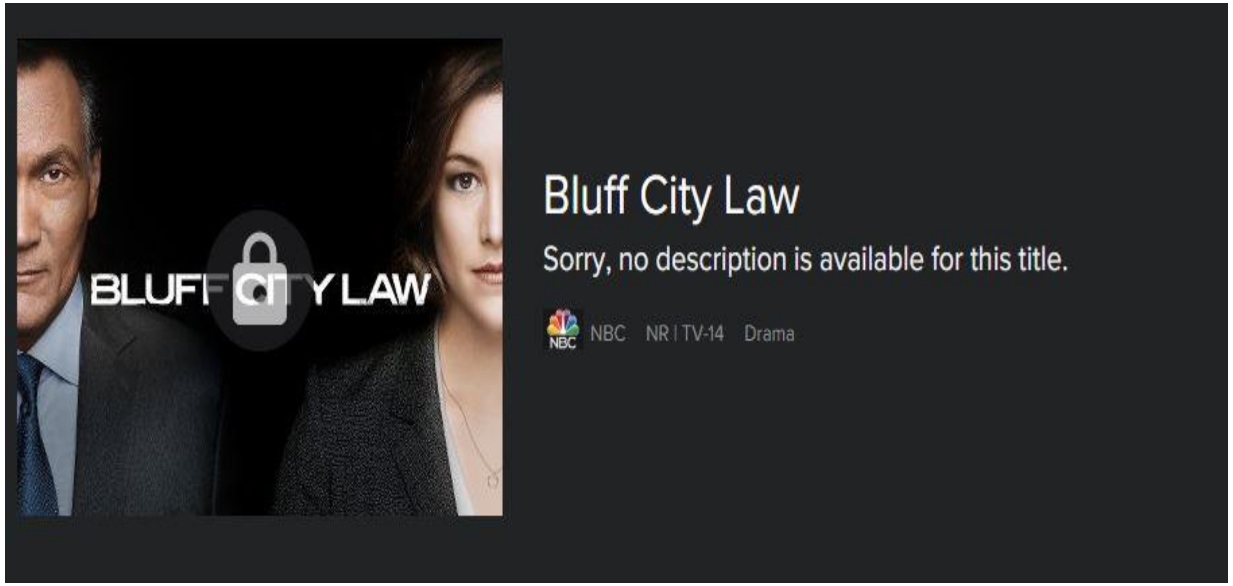
Example: Attribute – No Description

Historically: 17.3% are missing Description

Sample	Population	Percent	95% Cnf	99% Cnf
600	6,300	17.3%	2.88%	3.79%



95% No Descripiton



Bluff City Law

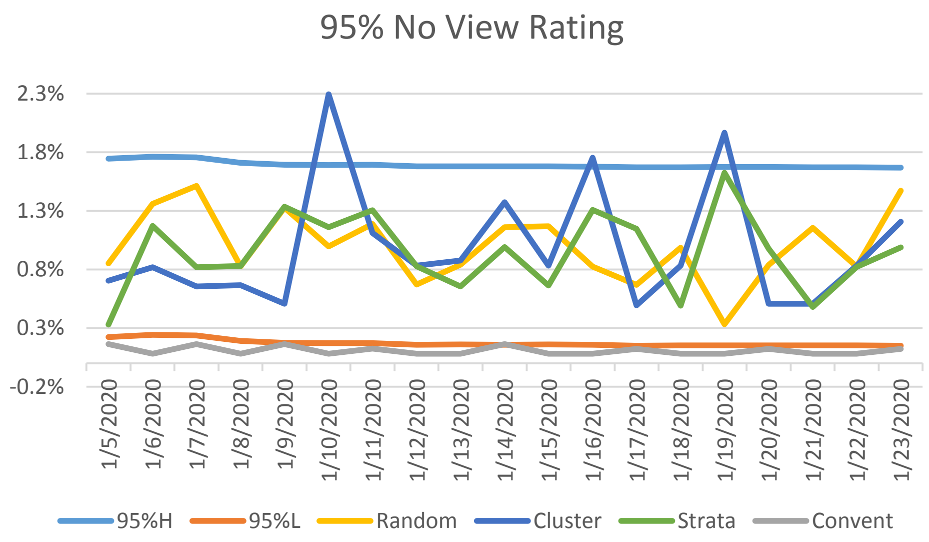
Sorry, no description is available for this title.

NBC NR | TV-14 Drama

Example: Attribute – No View Rating

Historically: 1.0% are missing View Rating

Sample	Population	Percent	95% Cnf	99% Cnf
600	6,300	1.0%	0.76%	1.00%



Long Live the Royals

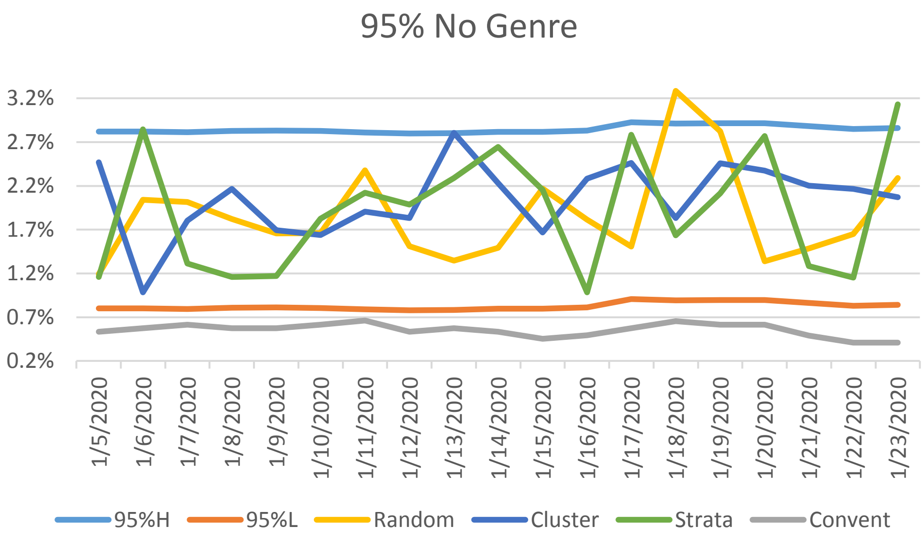
A British royal family balances ruling their kingdom with maintaining a normal family life as they honor the annual Yule Hare Festival.

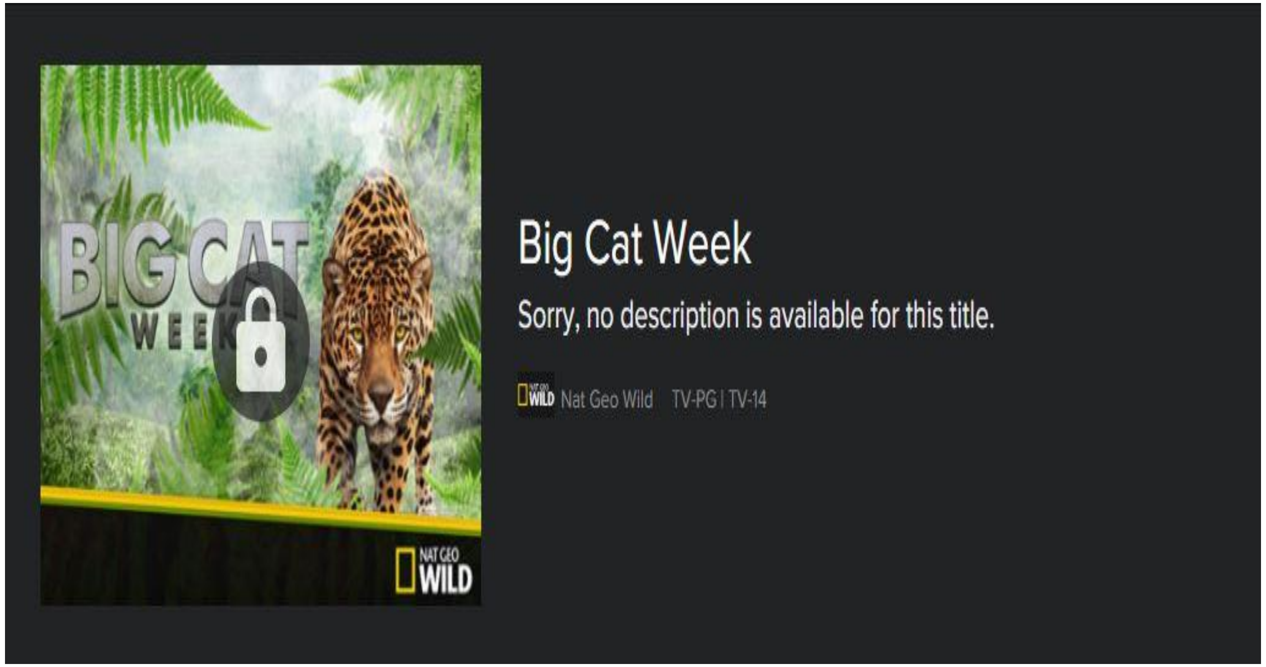
CN Cartoon Network Animated, Cartoon, Short Subject, Comedy, Family, Preteen

Example: Attribute – No Genre

Historically: 1.8% are missing Genre

Sample	Population	Percent	95% Cnf	99% Cnf
600	6,300	1.8%	1.01%	1.33%





Big Cat Week

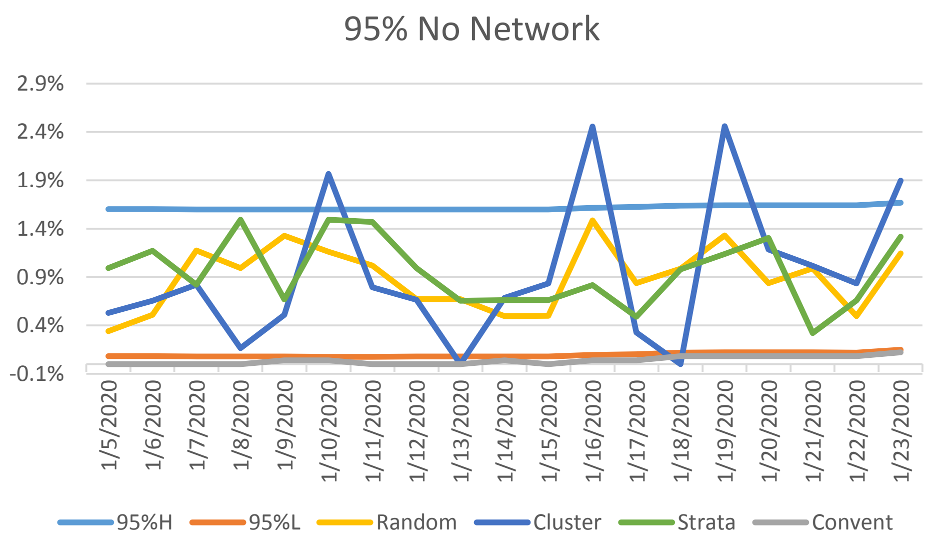
Sorry, no description is available for this title.

Nat Geo Wild TV-PG | TV-14

Example: Attribute – No Network

Historically: 1.0% are missing Network

Sample	Population	Percent	95% Cnf	99% Cnf
600	6,300	1.0%	0.76%	1.00%



Building Off the Grid

A newly married couple move off the grid to the Pacific Northwest.

Home & Garden, Home Projects, Outdoors

Cluster: Issue (19 Jan 2020)

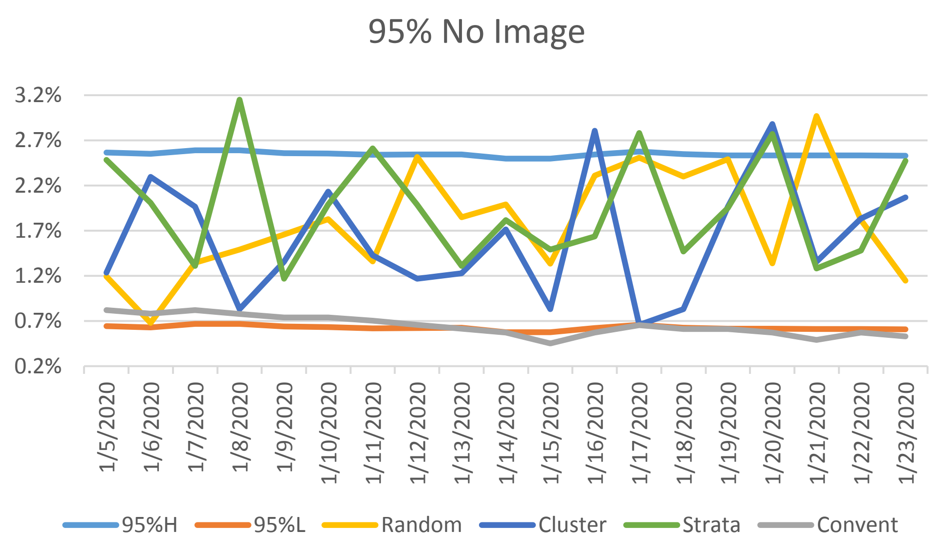
No Network: 3 same TV Show Titles have 13 of 15;
No Rating: 2 same TV Show Titles have 9 of 12;
No Image: 1 same TV Show Title has 4 of 12;

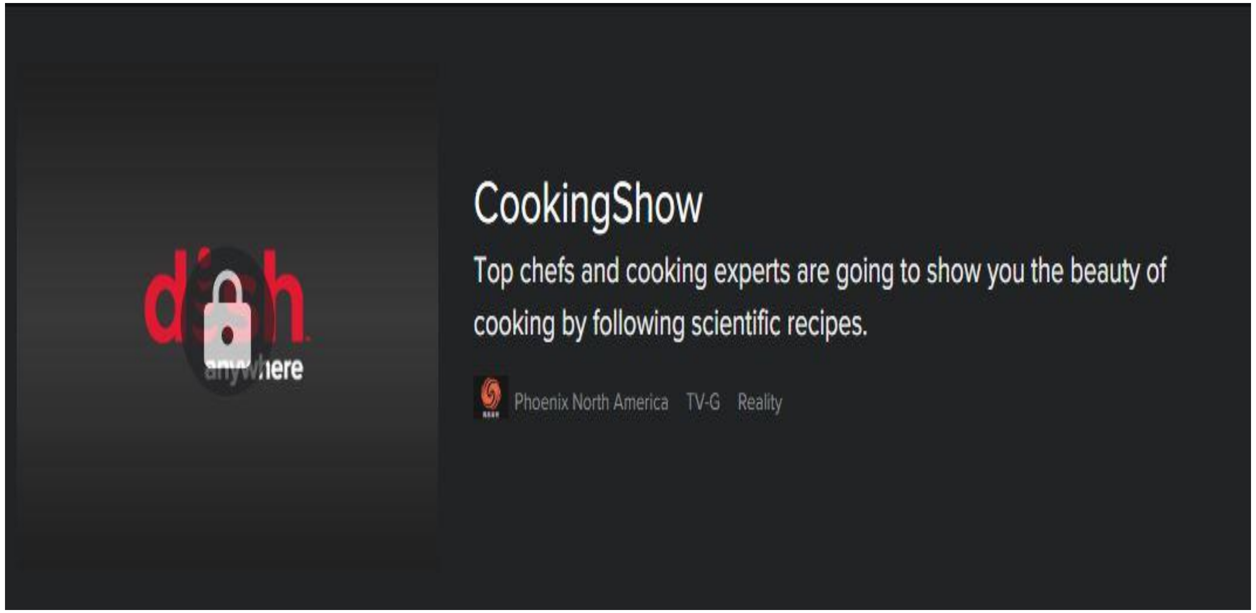
Missing	Franchise Id	TV Show Title
Network, Rating	building_off_the_grid_e2139260	Building Off the Grid
Network, Rating	building_off_the_grid_e2697018	Building Off the Grid
Network, Rating, Image	building_off_the_grid_e2793630	Building Off the Grid
Network, Rating	building_off_the_grid_e3006918	Building Off the Grid
Network, Desc, Rating, Image	building_off_the_grid_e3033997	Building Off the Grid
Network, Desc, Rating, Image	building_off_the_grid_e2939102	Building Off the Grid
Network, Image	building_off_the_grid_e3454375	Building Off the Grid
Network, Rating	building_off_the_grid_e2966006	Building Off the Grid

Example: Attribute – No Image

Historically: 1.6% are missing Image

Sample	Population	Percent	95% Cnf	99% Cnf
600	6,300	1.6%	0.96%	1.26%





CookingShow

Top chefs and cooking experts are going to show you the beauty of cooking by following scientific recipes.

Phoenix North America TV-G Reality

Conclusion and Thanks

Via this Compare and Analysis

- Prefer **Cluster** (sort: Title) 75% less time than **Simple Random**
  - More Cluster Analysis via Sort: Popular, Date added, Critics, Relevance
- Continue **Simple Random** or **Stratified**
  - When Cluster outside 95% limits
  - At least weekly to spot check **Cluster**

Thanks

- Arnold Miller, Sr. SDET (Sw. QA-Tester)
- Email: arnold.miller0@gmail.com
- Linkedin: http://www.linkedin.com/pub/arnold-miller/13/47b/a87