EMS ANALYTICS

Consumer Insight Consultants

Correlation and Regression for Drivers

Topics for Correlation and Regression

- A Simple Example (just 2 variables)
  - The Data
  - The Correlation
  - The (Simple) Regression
- Multiple Regression Example
  - The Data (again)
  - The Regression
  - The Outputs
- Some More Advanced Stuff
  - Multicollinearity
  - Shapley Regression
  - Decision Trees and Random Forest
  - Correlation Network
  - Partial correlations

For More Detail

https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-analysis/

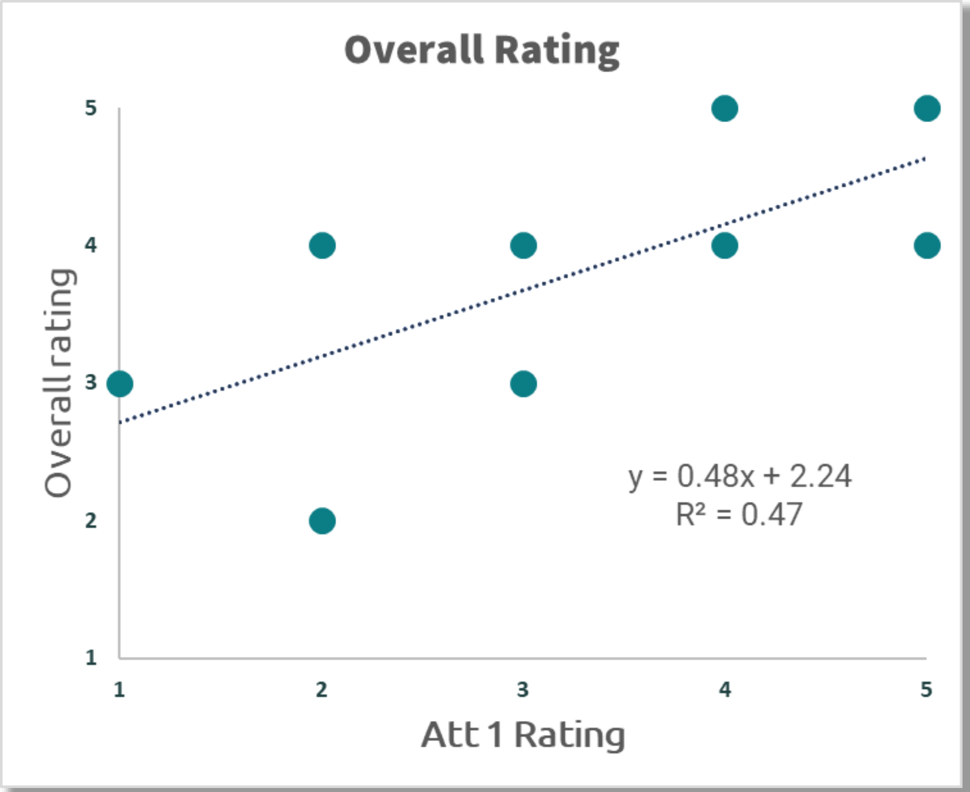
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/

What’s an easy way to see how one item “drives” another?

A Simple Example (just 2 variables)

The Data

Resp	Att 1 Rating	Att 2 Rating	Att 3 Rating	Overall Rating
1001	3	4	2	4
1002	4	2	5	4
1003	1	5	4	3
1004	5	4	3	5
1005	4	4	2	5
1006	5	3	2	3
1007	2	5	4	2
1008	3	4	3	4
1009	5	4	5	5



Custom Issues Require Custom Solutions

The Correlation

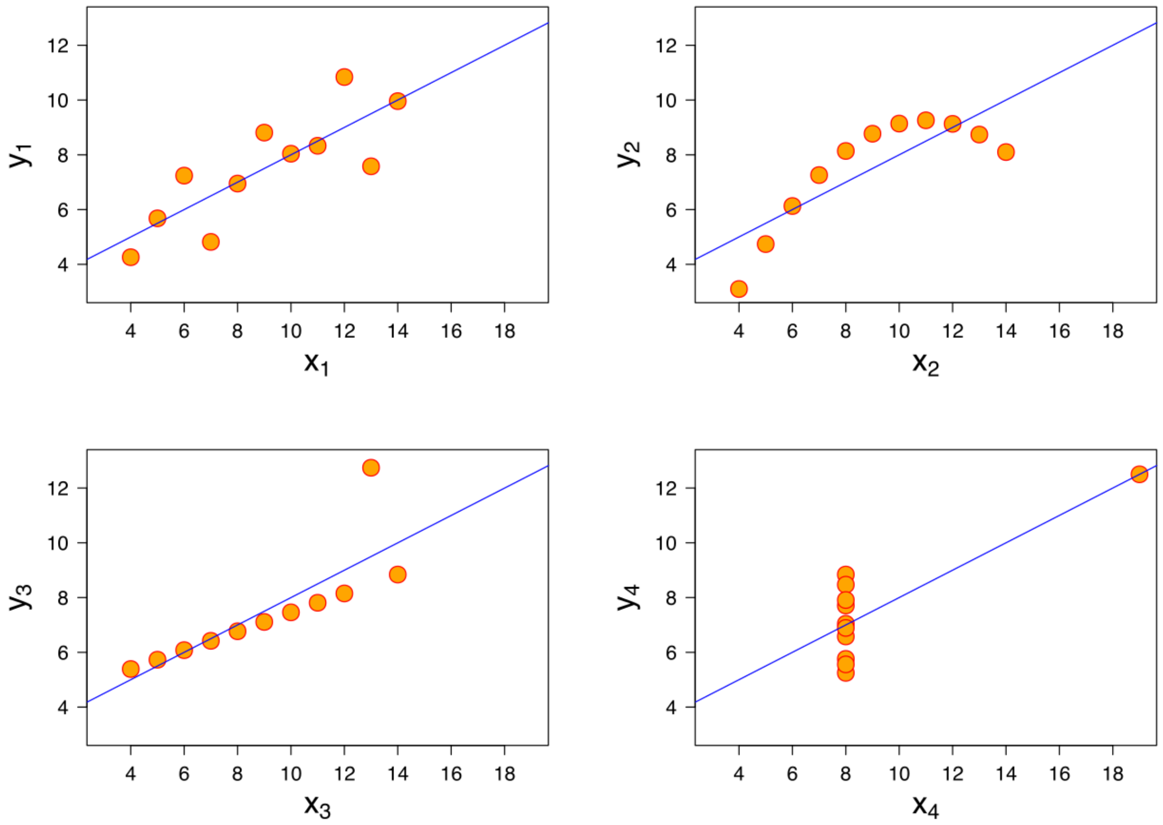
- A **correlation** is the strength of the linear relationship between two variables
  - +1.0 = perfect positive relationship
  - 0 = no relationship
  - -1.0 = perfect negative relationship
- For the previous data, the correlation (“r”) is about 0.69, which is usually considered moderately strong (depends on context)
- Correlations do not tell the whole story, but they are often used for driver analysis
- Beware spurious correlations and data that show “something else is going on” besides a linear effect

Correlations Often don't Tell the Whole Story

All four plots are for data with a 0.82 correlation.

Obviously (if it's plotted), there's something missing.

Moral: ALWAYS PLOT DATA before running correlations!



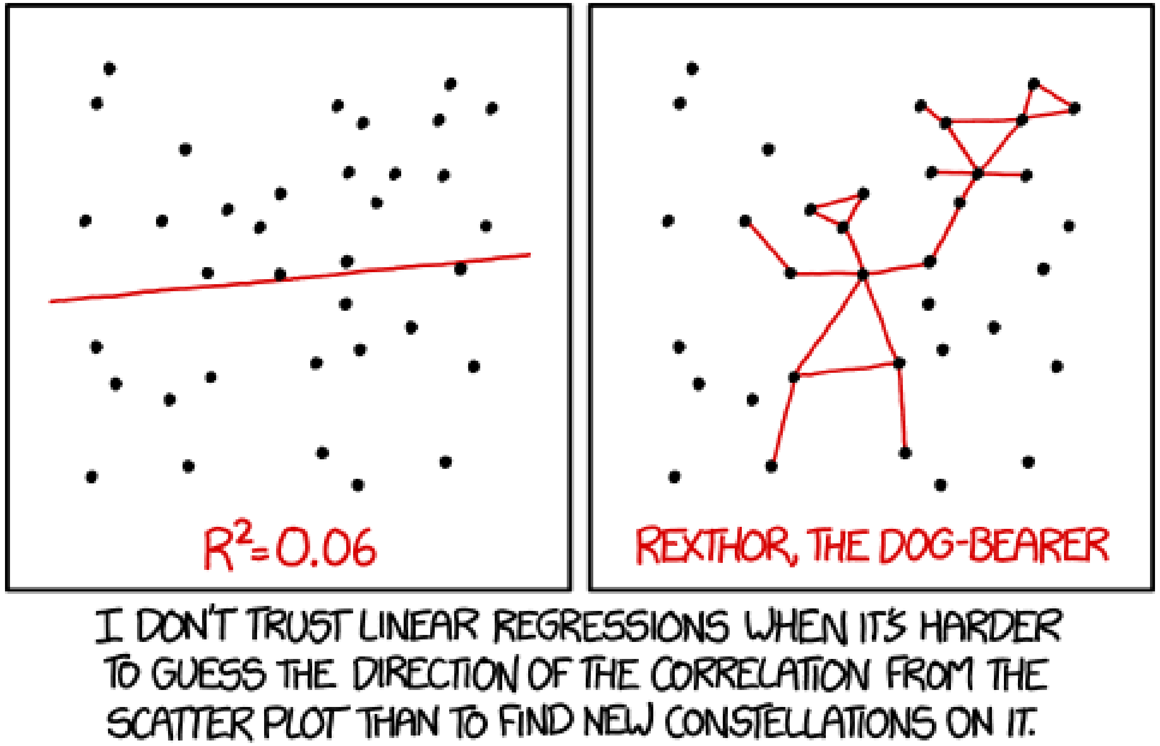
Final Notes on Correlations

- Actually, there are actually three “flavors” of correlation typically used:
  - Pearson – what we’ve been using so far and what’s used >95% of the time in MR
  - Spearman’s Rank – for rank data
  - Kendall’s Tau – also for rank data

The (Simple) Regression

- A (**simple**) **regression** is related in some ways to correlation, except it measures the magnitude of the linear relationship between two variables
  - Can be any value + or –
  - Goodness of fit is usually measured by the "r-squared", which is the percent variance of the target (dependent) variable explained by the other (independent) variable
  - standard error or the coefficient can also be used, but rarely is in MR
- For the previous data, the r-squared is about 0.47, which is usually considered moderately strong (depends on context)
- Like correlations, simple linear regressions do not tell the whole story, but they are often used for driver analysis

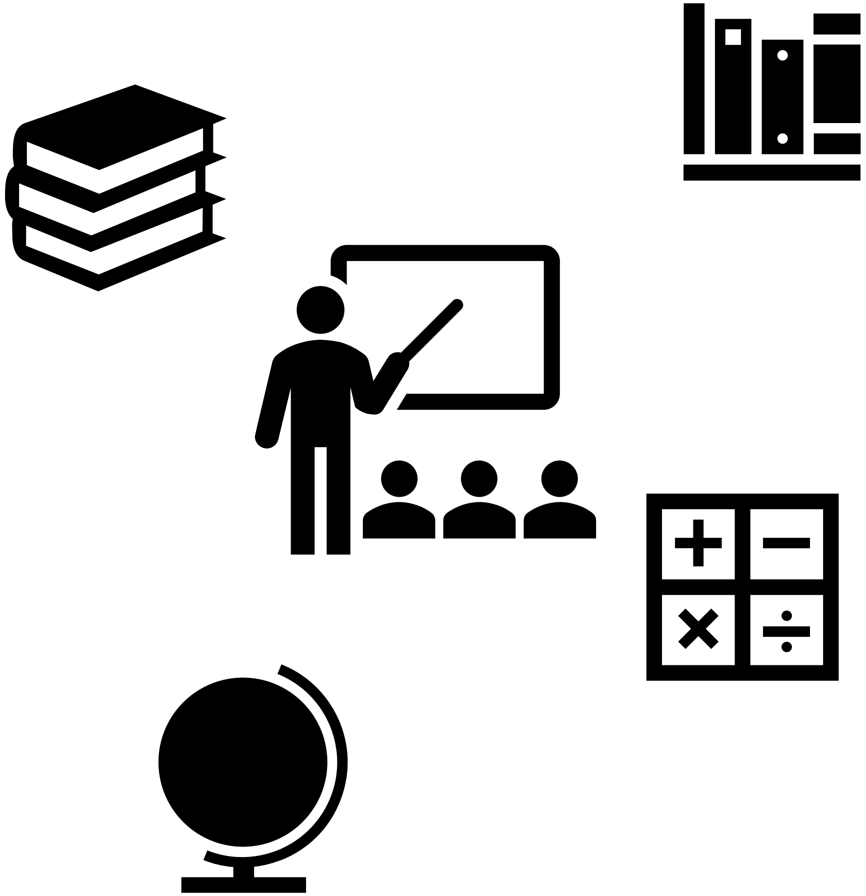
Regressions may not seem to (or actually) Tell the Whole Story



EMS ANALYTICS

Final Notes on (simple) Regressions

- It's not a coincidence – r-squared from regression is the square of the correlation ("r")
- While so far we have been talking about Ordinary Least Squares (OLS) in the form of $y=mx+b$, you may find that other forms are more appropriate
- Both correlations and regressions assume both variables are normally distributed and they can be heavily influenced by outliers
  - It's often necessary to standardize or at least transform the data (i.e., using the logarithm) to make the results reliable



What if I need to use regression on multiple variables at once?

Multiple Regression Example

The Data (same as before)

Resp	Att 1 Rating	Att 2 Rating	Att 3 Rating	Overall Rating
1001	3	4	2	4
1002	4	2	5	4
1003	1	5	4	3
1004	5	4	3	5
1005	4	4	2	5
1006	5	3	2	3
1007	2	5	4	2
1008	3	4	3	4
1009	5	4	5	5

The Multiple Regression

- Multiple regression is perhaps the most widely-used predictive tool used
- For the previous data, the r-squared for a single independent variable is about 0.47. If all three independent variables are used in the regression, we would expect to do even better.
  - In fact, R^2 is better with two more explanatory variables: 0.65
  - However, the adjusted R^2, which penalizes the fit for “extra” variables, is only 0.07
- The coefficients generally don’t mean anything specific and can be outright misleading

Multiple Regression Output

To the left we see the Excel Data Analysis output for a multiple regression on the same data we have been using.

Not shown are the correlations between the attributes of:

Att 1 & Att 2 – -0.61

Att 1 & Att 3 – -0.12

Att 2 & Att 3 – -.07

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.646
R Square	0.418
Adjusted R Square	0.069
Standard Error	1.017
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	3.715	1.238	1.197	0.400
Residual	5	5.174	1.035		
Total	8	8.889			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.964	3.195	0.302	0.775	-7.250	9.178	-7.250	9.178
Att 1 Rating	0.555	0.326	1.704	0.149	-0.282	1.392	-0.282	1.392
Att 2 Rating	0.216	0.498	0.434	0.682	-1.063	1.495	-1.063	1.495
Att 3 Rating	0.034	0.301	0.111	0.916	-0.740	0.807	-0.740	0.807

Fit Measures
Notice drop from R square to Adjusted R square. Std. Error is of the predictions

Coefficients
The b's (usually called betas) of the model

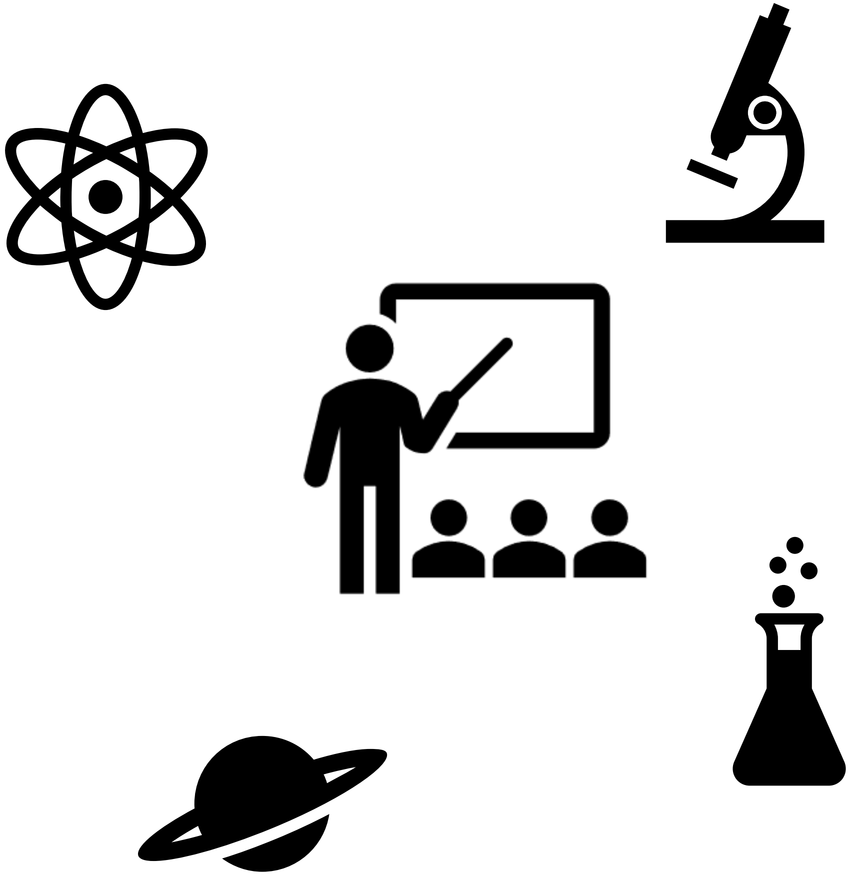
Model Significance
Is model better than no model?

Coeff. P-values
Is coefficient different from zero?

EMS ANALYTICS

Final Notes on Regressions

- The regression coefficients generally don't mean anything specific for driver analysis, but be sure to have the input data on the same scale and transform to a normal distribution if needed
- Beware of multicollinearity (correlated ind. Variables)
- Outliers can have a big impact on results, so clean them before running a regression.
- For driver analysis, p-values of coefficients are much more important than adjusted R^2
- There are other, better, ways to do drivers (we'll get to those in a few slides).



Some More Advanced Stuff

Beware Multicollinearity

- When the independent variables are correlated, which in survey research is almost always the case, that's called multicollinearity.
  - This can result in a lot of uncertainty in driver results
- Our example had a potential, even probable, multicollinearity problem with Att 1 and Att 2.
  - Since Att 2 and Att 3 are not correlated, Att 1 should be dropped from model, <u>but this doesn't work for drivers</u>!!
- Multicollinearity doesn't affect predictions, but it does affect coefficients

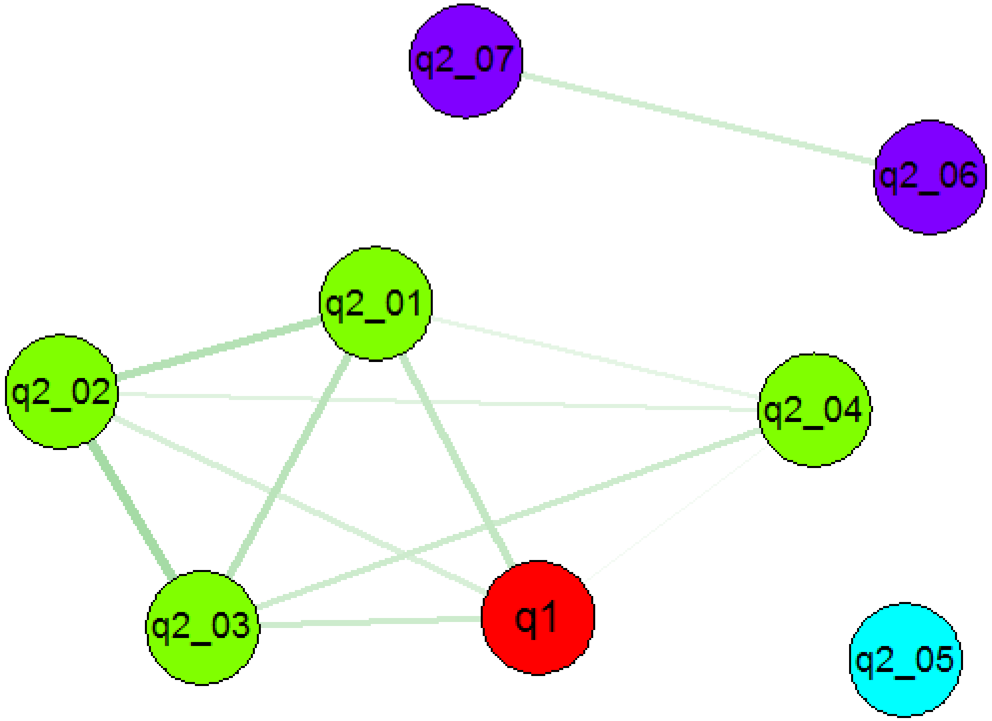
Shapley Regression

- One way to measure importance of attributes is to see which ones add the most to r^2 (explained variance) over all combinations of inputs.
- There are many names for this, but Shapley Regression is a common one.

Decision Trees and Random Forest

- Instead of OLS regression, decision trees (like CART) and their extension, random forest, can be used to estimate driver values.
- Like Shapley approach, these measure importance by an input's influence on model fit.
- Eliminates issues with the usual assumptions and doesn't have a problem with multicollinearity.
- Random forest is EMS's preferred method for drivers.

Correlation Network



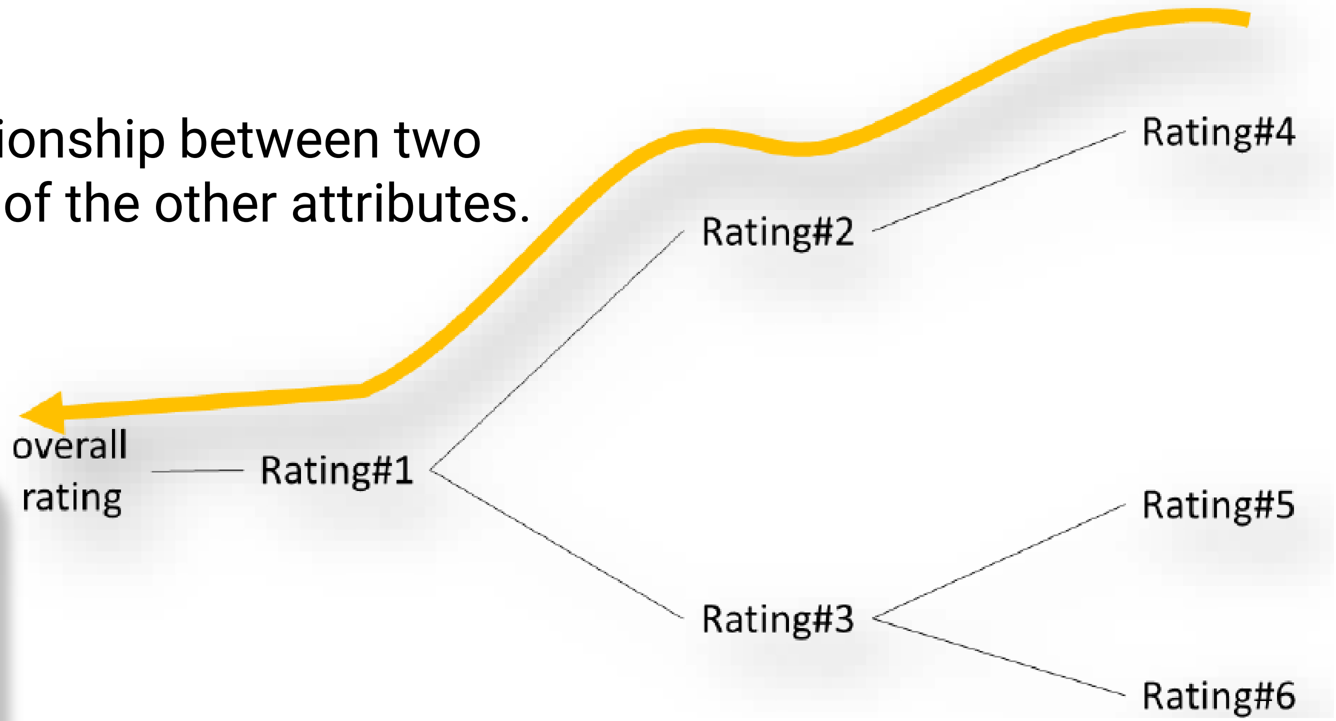
Design Thinking – Instead of listing values of drivers like they operate independently, use a correlation network to see how inputs interact.

In this example, q2_01 - 04 are all inter-related and 01 - 03 directly impact q1 (in this case, overall satisfaction). So, while q2_01 may be expensive to address, 02 and/or 03 may be better areas of focus.

Partial Correlation Network (Graphical Model)

Design Thinking – A great way to understand how to make improvements is with a partial correlation network.

Partial correlations are the relationship between two attributes, removing the effects of the other attributes.



Here, we see that ratings 4, 5, and 6 are good candidates for improving the overall rating. We would look to traditional drivers to help identify which ones are "best".

EMS ANALYTICS

Effective Solutions, Grounded Results