Association Analysis

a.k.a.

Market-Basket Analysis

Affinity Analysis

The Famous Example

- June 1992 Study by NCR (now TeraData) for Osco Drug searched for product associations
  - Beer and Diapers likely to be purchased together
  - (and many others, including Fruit Juice and Cough Syrup)

Association Analysis

- **Unsupervised**: No target outcome for training. Searching for patterns in the data.
- Association Analysis gives us sets of products that are likely to be purchased together.
- Used in retail for coupon marketing, targeted upselling, and product placement.
- Flexible analytics tool that can be used in many situations!

Association Rules

Data comes in the form of transactions:

Transaction ID	**Items**
10001	Bread, Juice
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, Bread, Formula
10004	Juice, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, Bread
10006	Formula, Juice, Diapers, Bread

Question: are there any relationships between items that might be hiding in these transactions?

Association Rules

Data comes in the form of transactions:

Transaction ID	Items
10001	Bread, Juice
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, Bread, Formula
10004	Juice, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, Bread
10006	Formula, Juice, Diapers, Bread

Diapers ➔ Formula?

An Association Rule

Diapers ➔ Formula

Antecedent

Consequent

Interpretation: Someone who buys diapers is also likely to (simultaneously) buy formula.

Quantifying Association Rules

- The strength of an association rule A ➔ B is quantified using **three statistics:**
  - **Support**: $P(A \cap B) = P(A \text{ and } B)$
    - Measures how often we find instances of this rule in the training data.
  - **Confidence**: $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$
    - Measures what percent of transactions containing A also contain B.
  - **Lift**: $\dfrac{P(B \mid A)}{P(B)} = \dfrac{P(A \cap B)}{P(A)P(B)}$
    - Measures how much more likely we are to buy B given that we also buy A than we are to buy B at random.
    - Want Lift values greater than 1 !!

Quantifying Association Rules

Transaction ID	Items
10001	Bread, Juice
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, Bread, Formula
10004	Juice, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, Bread
10006	Formula, Juice, Diapers, Bread

Diapers ➔ Formula?

Support: How often does this “rule” present itself?
P(Diapers and Formula)

In 50% of the transactions

Quantifying Association Rules

Transaction ID	Items
10001	Bread, Juice
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, Bread, Formula
10004	Juice, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, Bread
10006	Formula, Juice, Diapers, Bread

Diapers ➔ Formula?

Confidence: Given one buys Diapers, what is chance they also buy Formula? P(Formula | Diapers)

75% of Diaper purchases also contained Formula

Quantifying Association Rules

Transaction ID	Items
10001	Bread, Juice
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, Bread, Formula
10004	Juice, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, Bread
10006	Formula, Juice, Diapers, Bread

Diapers ➔ Formula?

Lift: How much more likely are we to see Formula with Diapers than see Formula overall? P(Formula|Diapers)/P(Formula)

We are (0.75/0.5 =) 1.5 times more likely to see a formula purchase with Diapers than we are to see one at random.

Some Post-Hoc Takeaways

Product A ➔ Product B

- Product B as consequent: Determine what can be done to boost its sales.
  - Product placement
  - Optimize upselling
  - Coupons for related products
- Product A as antecendent: Determine what other products would be affected by changes to Prod A.
  - Discontinuations
  - Price changes
  - Cannibalization

Direction of Association Rules

A→B vs. B→A

Same Support

Same Lift

Different Confidence

NO TIME COMPONENT

We DO NOT say "those who buy A will THEN buy B."

Finding Association Rules

- Most algorithms have two parts:
  - **Itemset generation** – find all sets of items that satisfy some minimum support.
  - **Rule generation** – determine which sets generated in step 1 satisfy some minimum confidence.
  - For more details of each part, see text by *Tan, Steinbach, and Kumar*
- To perform this analysis in any programming software, you need to specify:
  - **"The basket" -** something that identifies the basket of items being purchased simultaneously by a single customer.
  - **"The items" -** whatever column itemizes the transaction of each basket.
- The format for each software package/library will surely be different. Read the documentation!

SAS Viya Code

```
cas;
caslib _all_ assign;

proc mbanalysis data=public.grocery pctsupport=.5;
    customer transaction;
    target item;
    output out=casuser.setmb
           outfreq=casuser.freqmb
           outrule=casuser.mba_rules;
run;
```

SAS Viya Code

```
cas;
caslib _all_ assign;

proc mbanalysis data=public.grocery pctsupport=.5;
    customer transaction;#TheBasket”
    target item; #TheItems
    output out=casuser.setmb
           outfreq=casuser.freqmb
           outrule=casuser.mba_rules;
run;
```

You **must specify** *either of the following options:*

★ **SUPPORT**=*number*

specifies the minimum level of support (minimum frequency of an item) for a rule, where *number* must be an integer greater than or equal to 0. This option overrides the specification of the PCTSUP= option.

★ **PCTSUP**=*number*

★ **SUPPCT**=*number*

★ **SUP_PCT**=*number*

★ **PCTSUPPORT**=*number*

specifies the minimum level of support for a rule as a percentage of the number of baskets in the input data table, where *number* must be a real number between 0 and 100, inclusive. This option is ignored if the SUPPORT= option is specified.

You **can also specify** the following *options*:

★ **CONF**=*number*

specifies the minimum confidence for the rules, where *number* must be a real number between 0 and 100.

By default, CONF=50

★ **ITEMS**=*number*

specifies the number of items in a rule, where *number* must be an integer between 1 and 100.
By default, ITEMS=2 when either an OUT= or OUTRULE= option is specified in the OUTPUT statement; otherwise, ITEMS=1 by default.

★ **LIFT**=*number*

specifies the minimum lift value necessary to generate a rule, where *number* must be a positive, real number between 0 and 100, inclusive. By default, LIFT=1.

Viya Output

MBA_RULES Output Dataset

RULEID	LHS	RHS	COUNT	SUPPORT	CONF	LIFT	ITEM1	ITEM2	RULE
9	1	1	62	0.630	51.240	2.129	brown	whole milk	brown ==> whole milk
16	1	1	50	0.508	51.546	6.243	other	citrus fruit	other ==> citrus fruit
20	1	1	52	0.529	56.522	3.090	cream	other vegetables	cream ==> other vegetables
28	1	1	69	0.702	54.762	2.275	frozen	whole milk	frozen ==> whole milk
37	1	1	66	0.671	51.163	2.126	newspa	whole milk	newspa ==> whole milk
40	1	1	98	0.996	53.846	2.944	rolls/	other vegetables	rolls/ ==> other vegetables
41	1	1	66	0.671	60.000	3.280	whippe	other vegetables	whippe ==> other vegetables
42	1	1	56	0.569	57.732	5.465	other	root vegetables	other ==> root vegetables

Unique Rule Identifier

Viya Output

MBA_RULES Output Dataset

RULEID	LHS	RHS	COUNT	SUPPORT	CONF	LIFT	ITEM1	ITEM2	RULE
9	1	1	62	0.630	51.240	2.129	brown	whole milk	brown ==> whole milk
16	1	1	50	0.508	51.546	6.243	other	citrus fruit	other ==> citrus fruit
20	1	1	52	0.529	56.522	3.090	cream	other vegetables	cream ==> other vegetables
28	1	1	69	0.702	54.762	2.275	frozen	whole milk	frozen ==> whole milk
37	1	1	66	0.671	51.163	2.126	newspa	whole milk	newspa ==> whole milk
40	1	1	98	0.996	53.846	2.944	rolls/	other vegetables	rolls/ ==> other vegetables
41	1	1	66	0.671	60.000	3.280	whippe	other vegetables	whippe ==> other vegetables
42	1	1	56	0.569	57.732	5.465	other	root vegetables	other ==> root vegetables

Number of items on the Left Hand Side (LHS) of the rule and the RHS

Viya Output

MBA_RULES Output Dataset

RULEID	LHS	RHS	COUNT	SUPPORT	CONF	LIFT	ITEM1	ITEM2	RULE
9	1	1	62	0.630	51.240	2.129	brown	whole milk	brown ==> whole milk
16	1	1	50	0.508	51.546	6.243	other	citrus fruit	other ==> citrus fruit
20	1	1	52	0.529	56.522	3.090	cream	other vegetables	cream ==> other vegetables
28	1	1	69	0.702	54.762	2.275	frozen	whole milk	frozen ==> whole milk
37	1	1	66	0.671	51.163	2.126	newspa	whole milk	newspa ==> whole milk
40	1	1	98	0.996	53.846	2.944	rolls/	other vegetables	rolls/ ==> other vegetables
41	1	1	66	0.671	60.000	3.280	whippe	other vegetables	whippe ==> other vegetables
42	1	1	56	0.569	57.732	5.465	other	root vegetables	other ==> root vegetables

Number of grocery baskets in which the rule appears

Viya Output

MBA_RULES Output Dataset

RULEID	LHS	RHS	COUNT	SUPPORT	CONF	LIFT	ITEM1	ITEM2	RULE
9	1	1	62	0.630	51.240	2.129	brown	whole milk	brown ==> whole milk
16	1	1	50	0.508	51.546	6.243	other	citrus fruit	other ==> citrus fruit
20	1	1	52	0.529	56.522	3.090	cream	other vegetables	cream ==> other vegetables
28	1	1	69	0.702	54.762	2.275	frozen	whole milk	frozen ==> whole milk
37	1	1	66	0.671	51.163	2.126	newspa	whole milk	newspa ==> whole milk
40	1	1	98	0.996	53.846	2.944	rolls/	other vegetables	rolls/ ==> other vegetables
41	1	1	66	0.671	60.000	3.280	whippe	other vegetables	whippe ==> other vegetables
42	1	1	56	0.569	57.732	5.465	other	root vegetables	other ==> root vegetables

Support, Confidence and Lift of the Rule

Viya Output

MBA_RULES Output Dataset

RULEID	LHS	RHS	COUNT	SUPPORT	CONF	LIFT	ITEM1	ITEM2	RULE
9	1	1	62	0.630	51.240	2.129	brown	whole milk	brown ==> whole milk
16	1	1	50	0.508	51.546	6.243	other	citrus fruit	other ==> citrus fruit
20	1	1	52	0.529	56.522	3.090	cream	other vegetables	cream ==> other vegetables
28	1	1	69	0.702	54.762	2.275	frozen	whole milk	frozen ==> whole milk
37	1	1	66	0.671	51.163	2.126	newspa	whole milk	newspa ==> whole milk
40	1	1	98	0.996	53.846	2.944	rolls/	other vegetables	rolls/ ==> other vegetables
41	1	1	66	0.671	60.000	3.280	whippe	other vegetables	whippe ==> other vegetables
42	1	1	56	0.569	57.732	5.465	other	root vegetables	other ==> root vegetables

All items involved in the rule are listed in separate columns.

Network Visualization

• • •

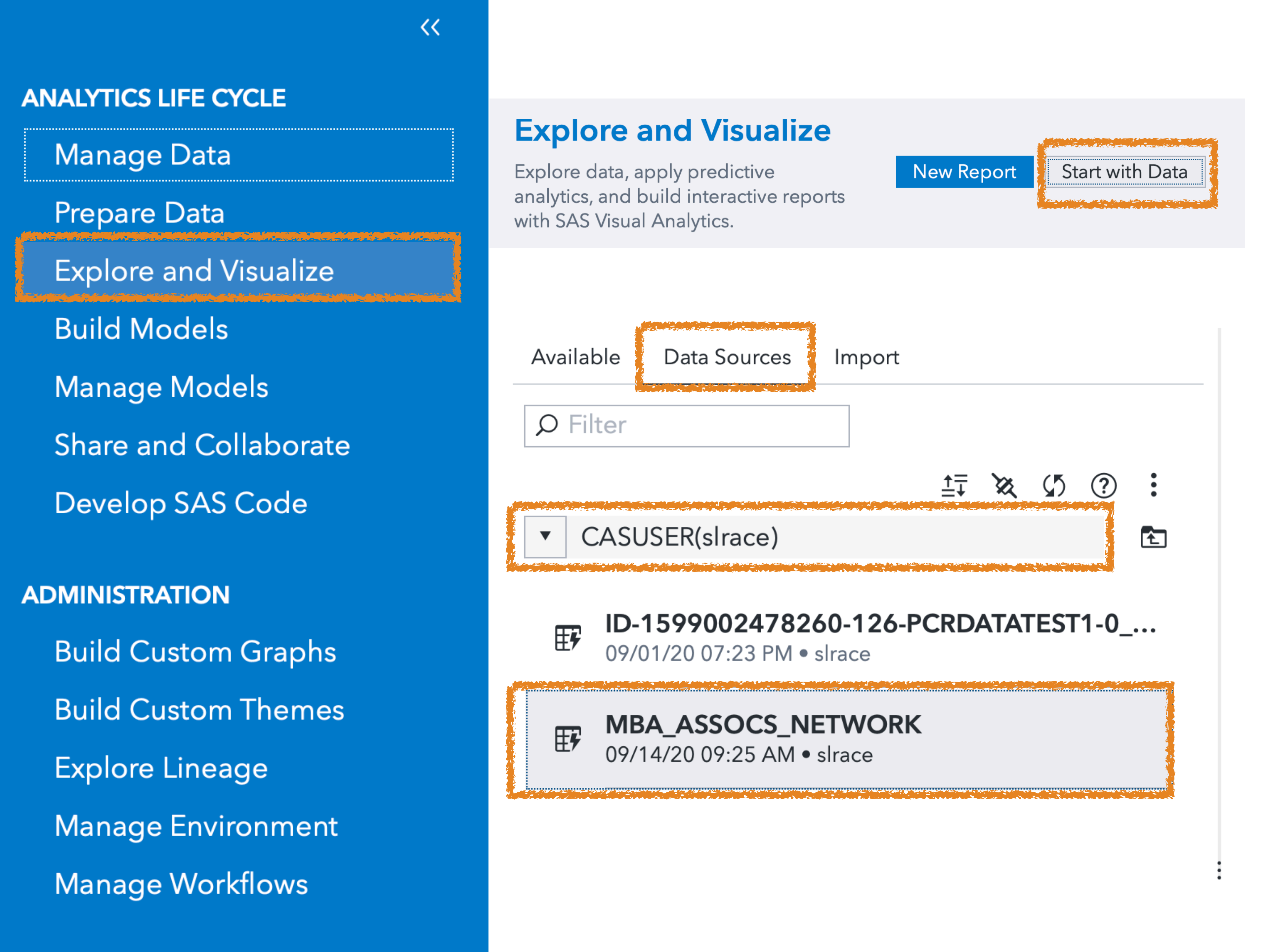
```
proc fedsql sessref = casauto;
    create table casuser.mba_network as
        select t1.item as t1_item,
            t1.count as item_count,
            t1.support as item_support,
            t2.*
    from casuser.freqmb as t1 inner join casuser.mba_rules as t2
        on t1.item=t2.item1;
    create table casuser.mba_network2 as
        select t1.item as t1_item,
            t1.count as item_count,
            t1.support as item_support
    from casuser.freqmb as t1 inner join casuser.mba_rules as t2
        on t1.item=t2.item2;
quit;

data casuser.mba_network_final;
    set casuser.mba_network casuser.mba_network2;
run;
```

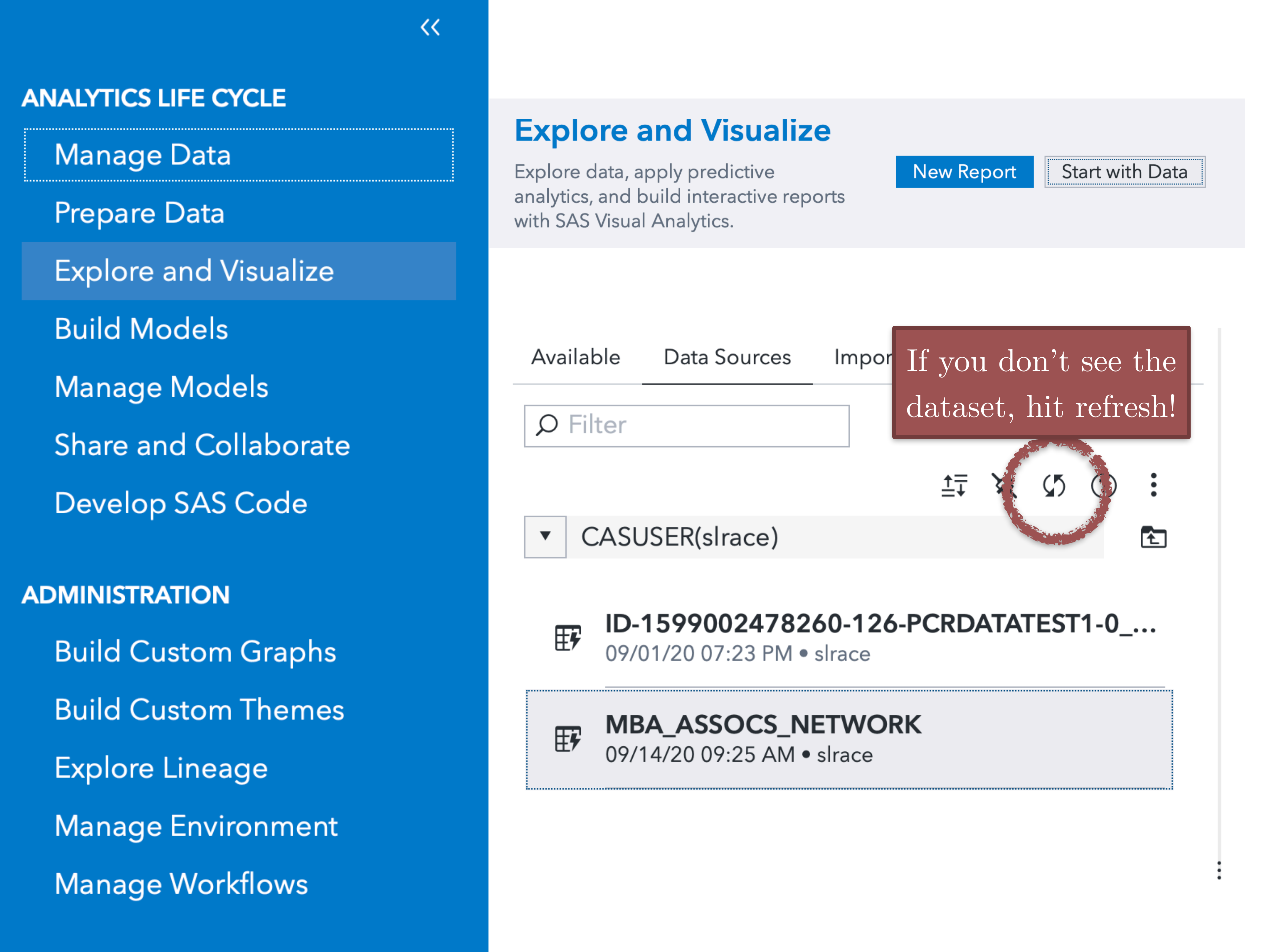
Create Network Dataset from Output

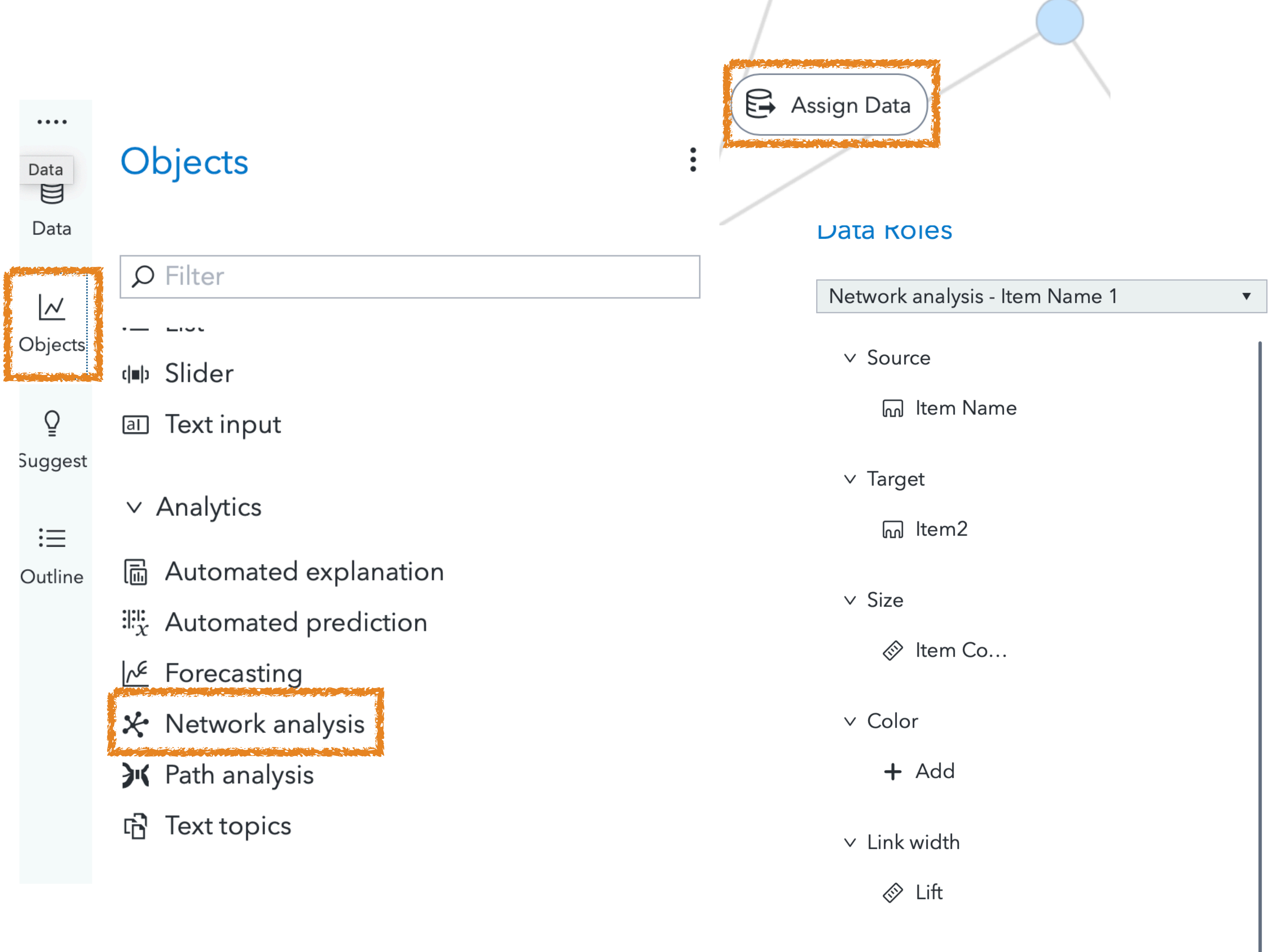
```
proc casutil;
    promote casdata='mba_network_final'
    incaslib ='casuser'
    outcaslib='casuser'
    casout ='mba_assocs_network';
run;
```

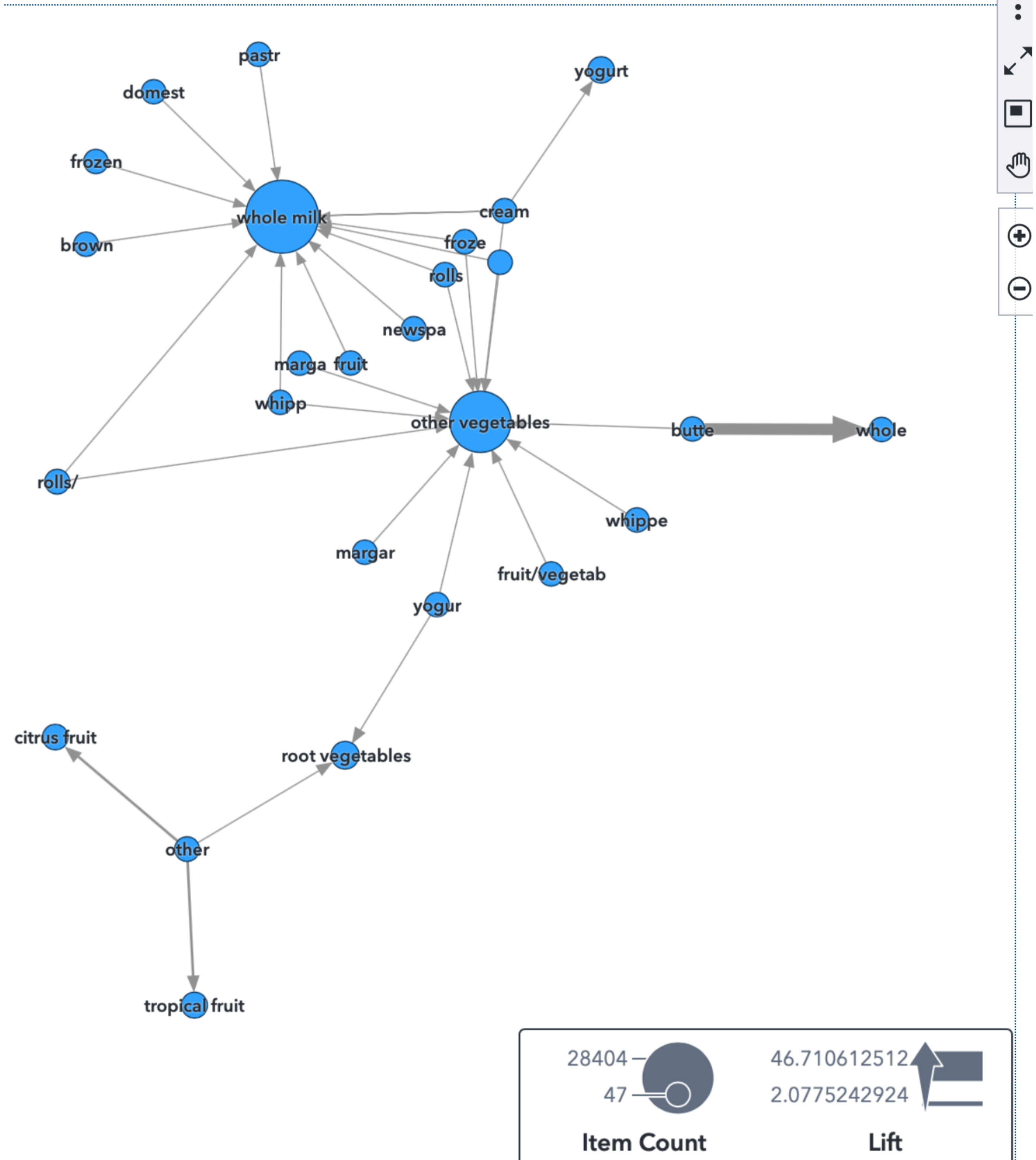
Promote Dataset for Visual Studio



ANALYTICS LIFE CYCLE
Manage Data
Prepare Data
Explore and Visualize
Build Models
Manage Models
Share and Collaborate
Develop SAS Code
ADMINISTRATION
Build Custom Graphs
Build Custom Themes
Explore Lineage
Manage Environment
Manage Workflows
Explore and Visualize
Explore data, apply predictive analytics, and build interactive reports with SAS Visual Analytics.
New Report
Start with Data
Available
Data Sources
Import
Filter
CASUSER(slrace)
ID-1599002478260-126-PCRDATATEST1-0_...
09/01/20 07:23 PM • slrace
MBA_ASSOCS_NETWORK
09/14/20 09:25 AM • slrace


ANALYTICS LIFE CYCLE
Manage Data
Prepare Data
Explore and Visualize
Build Models
Manage Models
Share and Collaborate
Develop SAS Code
ADMINISTRATION
Build Custom Graphs
Build Custom Themes
Explore Lineage
Manage Environment
Manage Workflows
Explore and Visualize
Explore data, apply predictive analytics, and build interactive reports with SAS Visual Analytics.
New Report
Start with Data
Available
Data Sources
Impor
If you don't see the dataset, hit refresh!
Filter
CASUSER(slrace)
ID-1599002478260-126-PCRDATATEST1-0_...
09/01/20 07:23 PM • slrace
MBA_ASSOCS_NETWORK
09/14/20 09:25 AM • slrace


Assign Data
Data
Data
Objects
Objects
Suggest
Outline
Filter
Slider
Text input
Analytics
Automated explanation
Automated prediction
Forecasting
Network analysis
Path analysis
Text topics
Data Roles
Network analysis - Item Name 1
Source
Item Name
Target
Item2
Size
Item Co...
Color
Add
Link width
Lift



pastr
domest
yogurt
frozen
whole milk
cream
froze
brown
rolls
newspa
marga
fruit
whipp
other vegetables
butte
whole
rolls/
whippe
margar
fruit/vegetab
yogur
citrus fruit
root vegetables
other
tropical fruit
28404
47
Item Count
46.710612512
2.0775242924
Lift