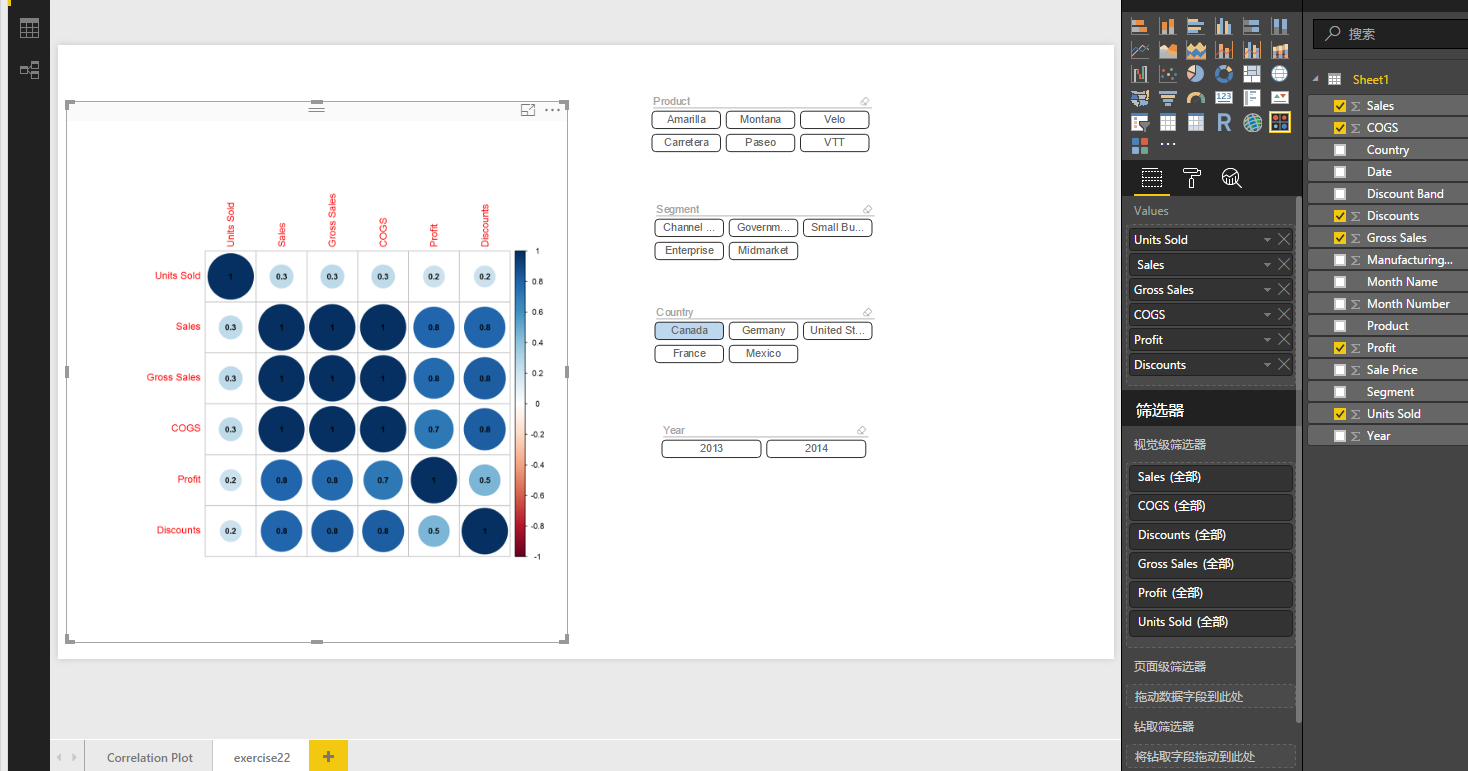
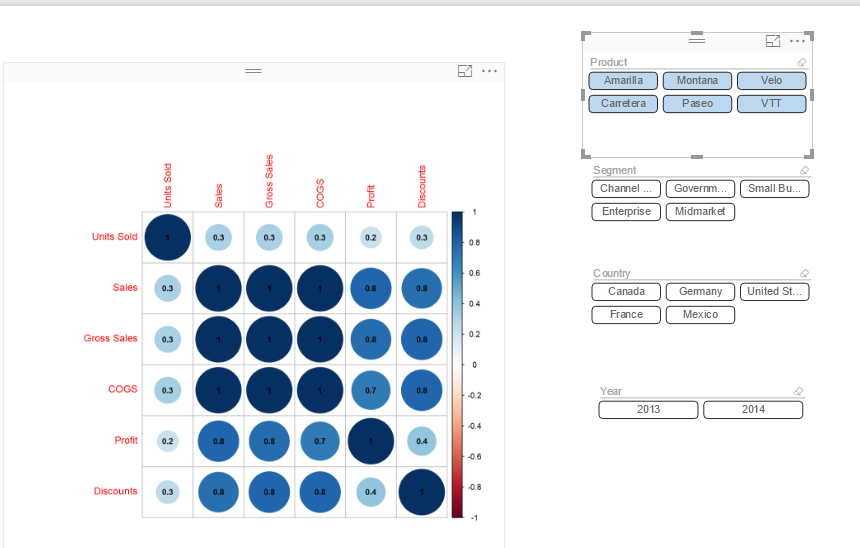
21. Once the report is created in Power BI Desktop, you can publish the report containing one or more R visuals to the Power BI service, R visuals currently can only be created in Power BI Desktop, and then published to the Power BI service, you can interact, filter, slice and pin them to a dashboard, or share them with other visuals, one difference from other visuals is that R visuals cannot show tool tips and cannot be used to filter other visuals.

22.



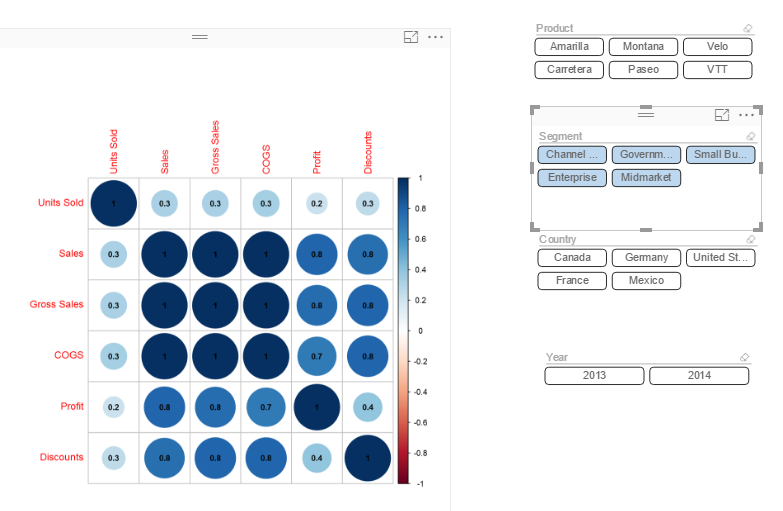
1. Does Discount correlate within Profit in all Product:

Yes, for all products, we can find there is some correlate between profit and discount, the value is 0.4



1. Does Discount correlate within Profit in all Segment?

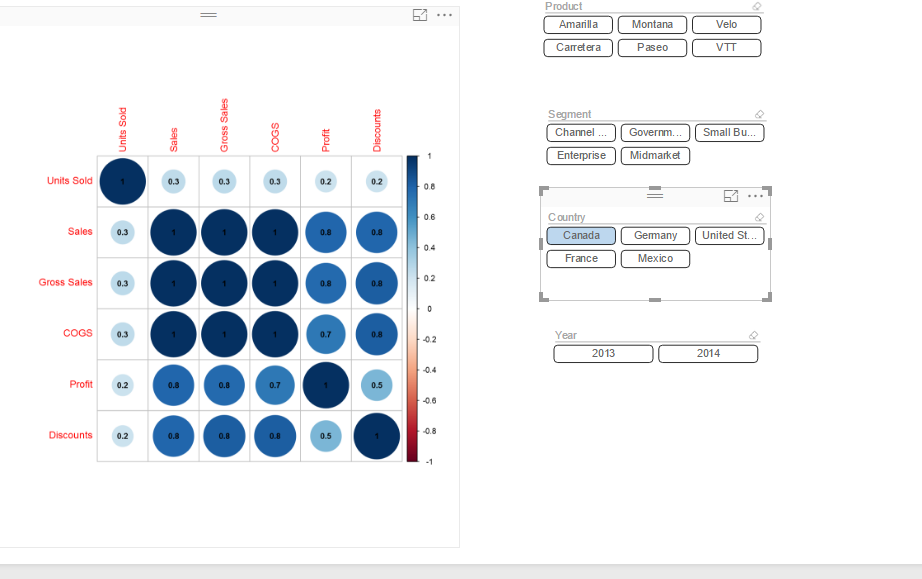
Yes, for all segments, we can find there is some correlation between discount and product, the value is 0.4



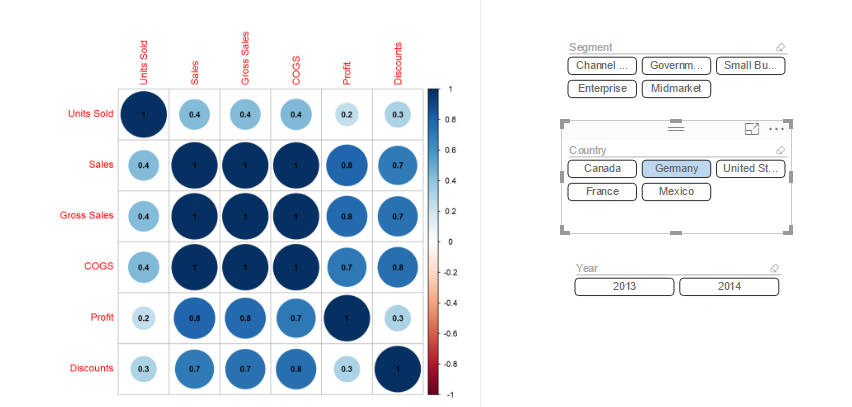
1. Are there differences by Country?

Yes, different country has different correlation values on these items such as COGS and Discount

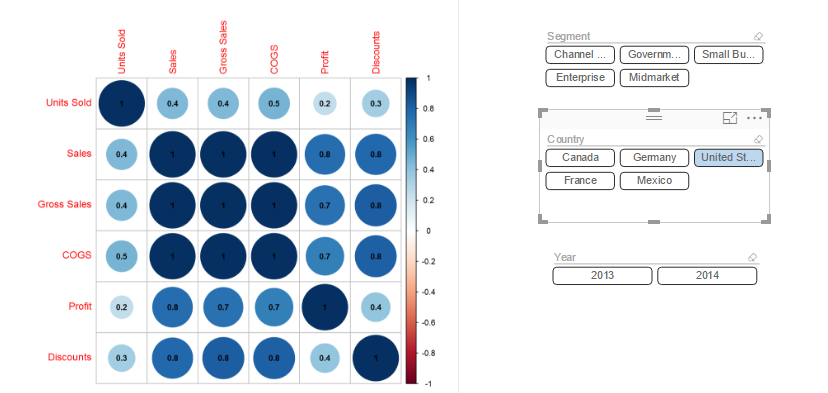
For Canada:



For Germany:



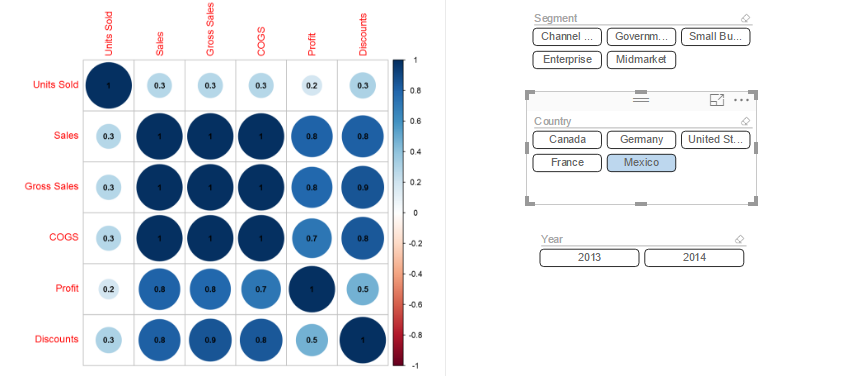
For USA:

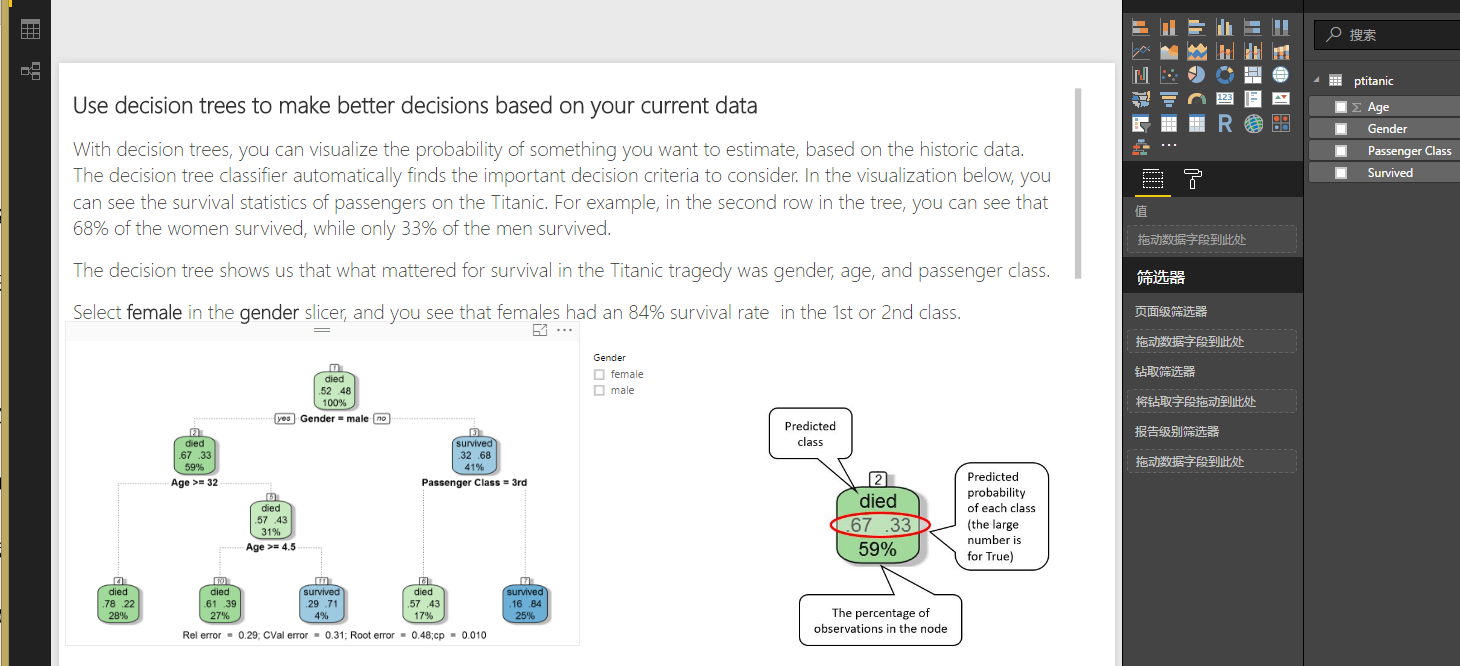


For France:



For Maxico:



23. 

24.

Accuracy:

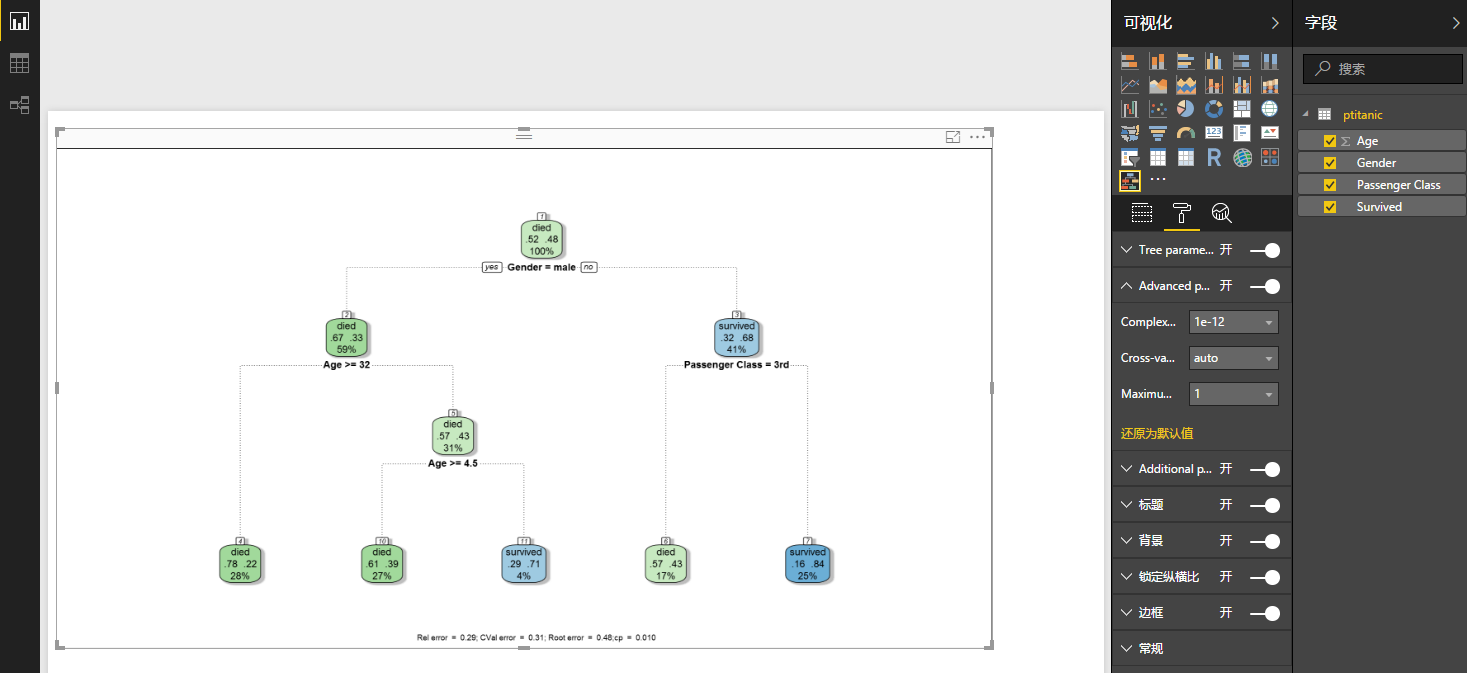
Precision:

Recall or Sensitivity:

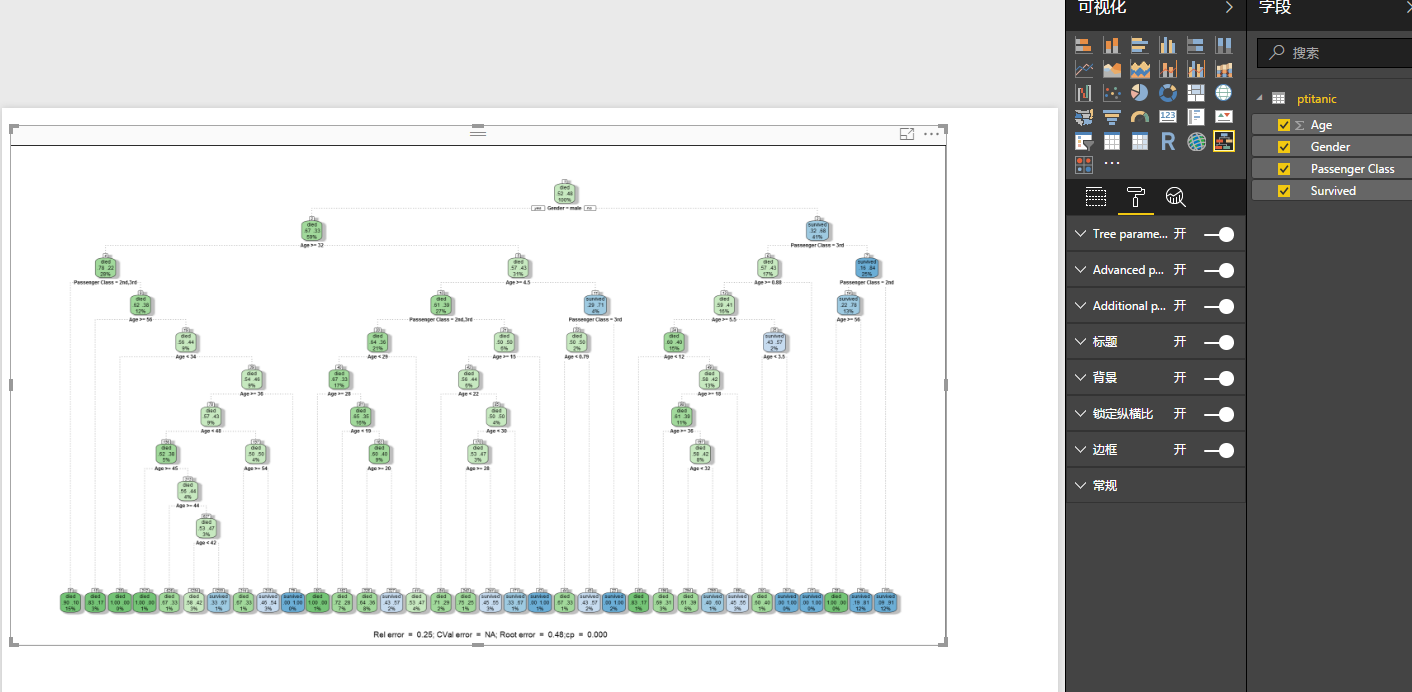
Specificity:

25.

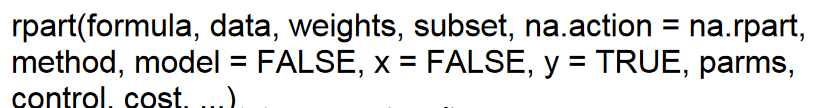
When advanced parameter’s cross-validation is auto



When advanced parameter’s cross-validation is ‘None’



Use rpart to build the decision tree:



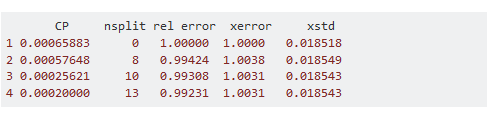
Then use test data to observe the accuracy.

Rel error: the relative error for predictions.

We can use the following parameters for decision tree as an example:

tree\_model <- rpart(Failed ~ race + gender + age+ time\_in\_hospital+ medical\_specialty + num\_lab\_procedures+ num\_procedures+num\_medications+number\_outpatient+number\_emergency+number\_inpatient+number\_diagnoses+max\_glu\_serum+ A1Cresult+metformin+glimepiride+glipizide+glyburide+pioglitazone+rosiglitazone+insulin+change,method="class", data=training\_data, control=rpart.control(minsplit=2, cp=0.0001, maxdepth=20, xval = 10), parms = list(split = "gini")

xprinting the results yields:

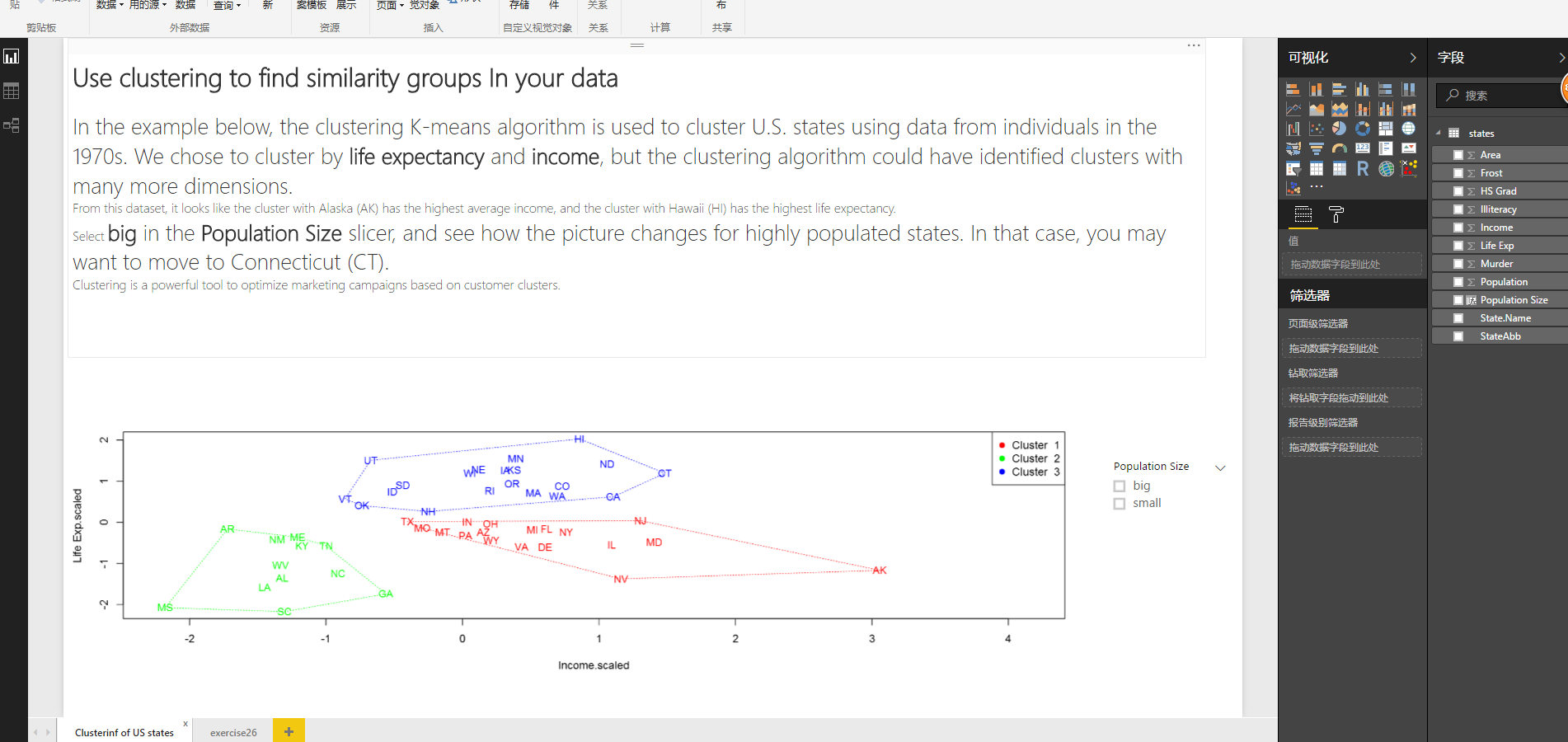


So the Rel error can be calculated by rpart function.

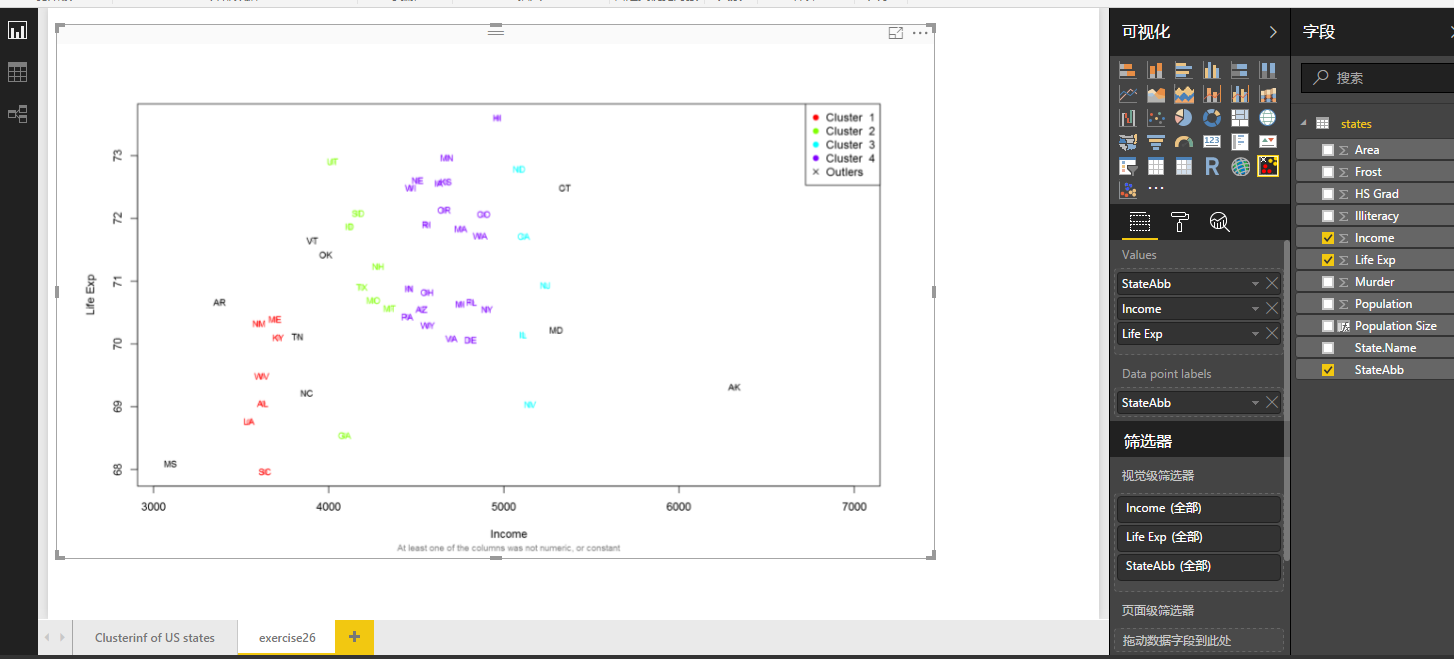
Root error: it is used to compute two measures of predictive performance, when considering values displayed in the rel error and xerror column, and depending on the complexity parameter.

Cp : The complexity parameter (cp) is used to control the size of the decision tree and to select the optimal tree size. The result can be calculated by rpart and result is in the first column of the figure above.

26. The imported file is in figure below:

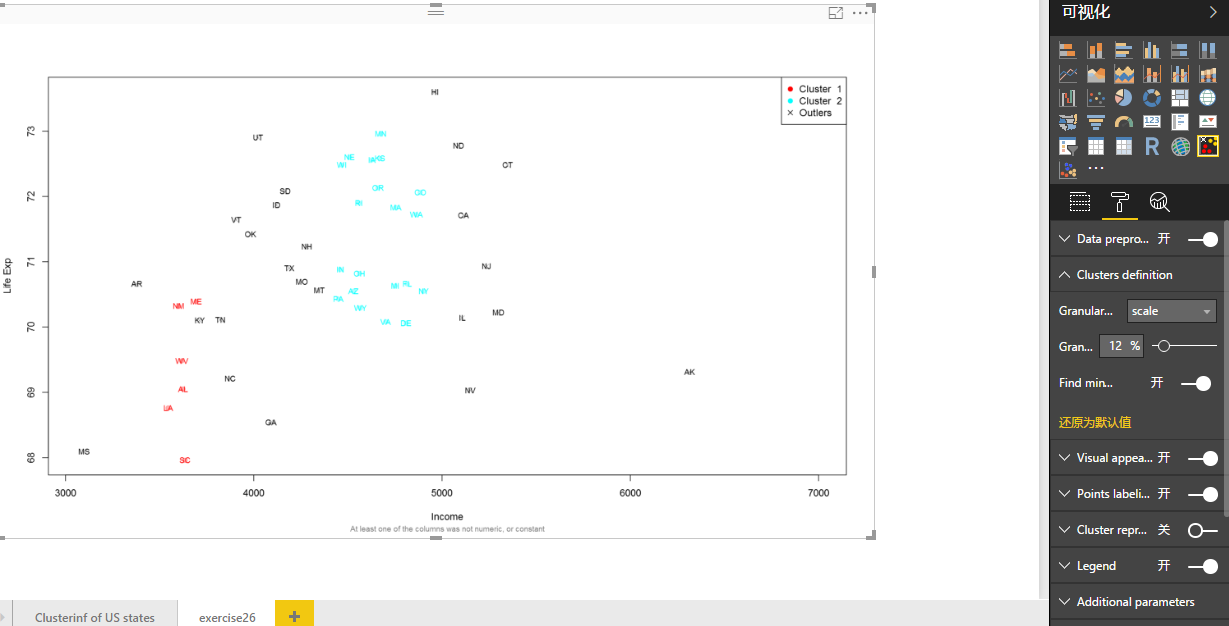


Inserted new page and make clusters (apply format setting Data preprocessing, Clusters definition and Visual appearance)

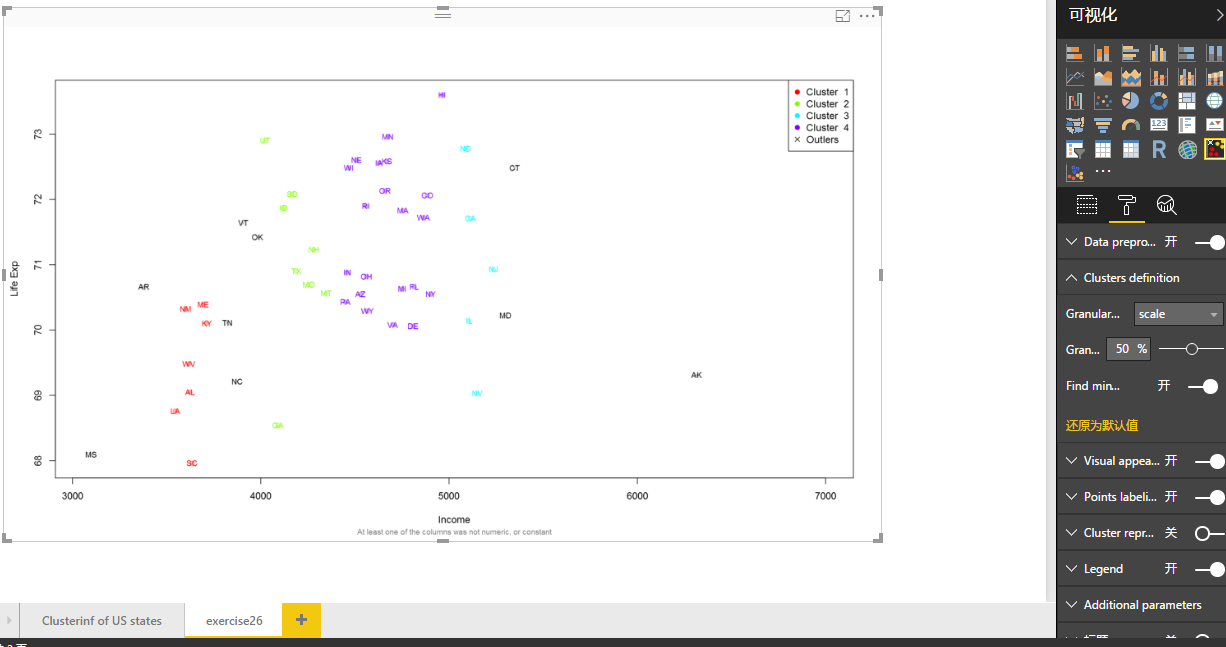




27. when applying the format, and change the granularity, when I choose granularity method: scale, and apply 12%, there are 2 clusters



When I apply 50% as the granularity, there are 4 clusters,

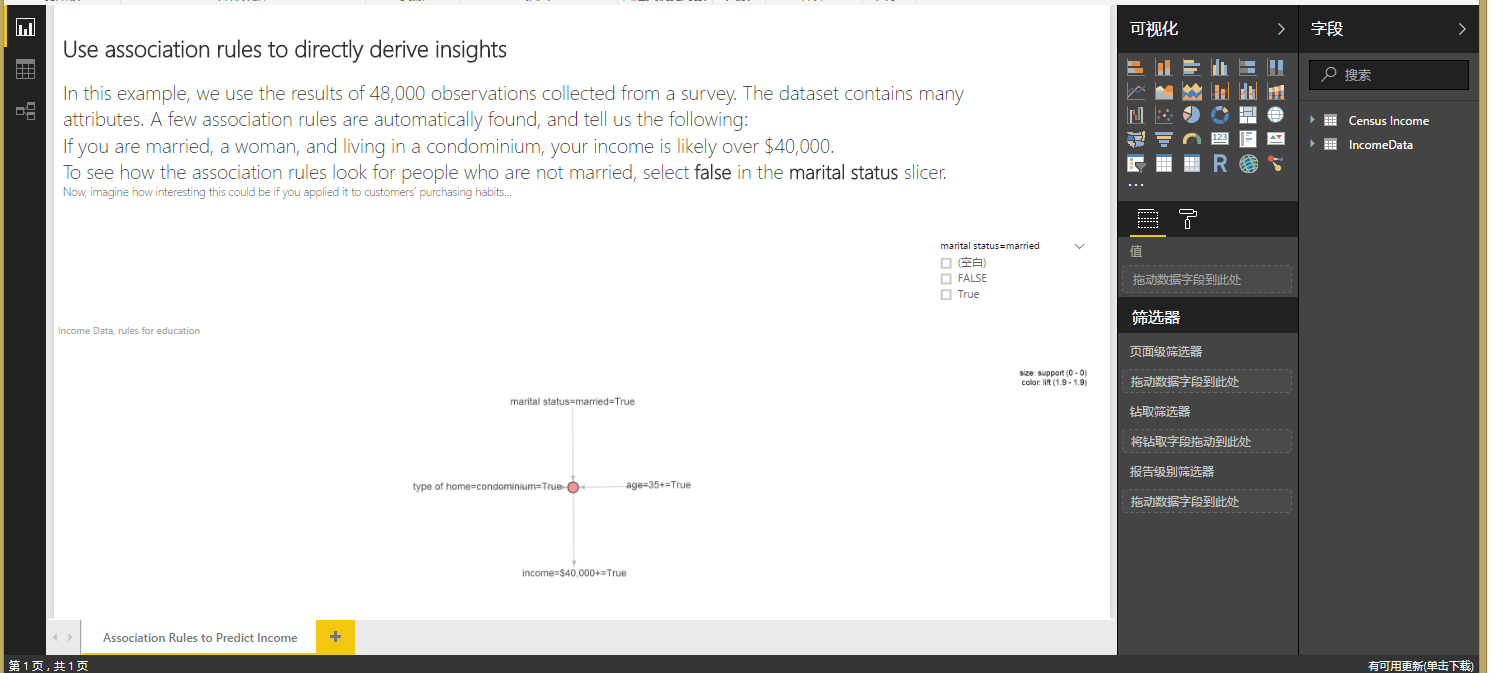


When I apply 80%, there is only 1 cluster.

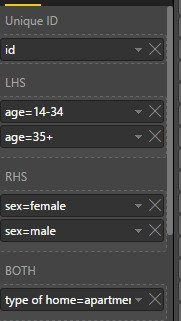
When I apply PCA, the result is in figure below



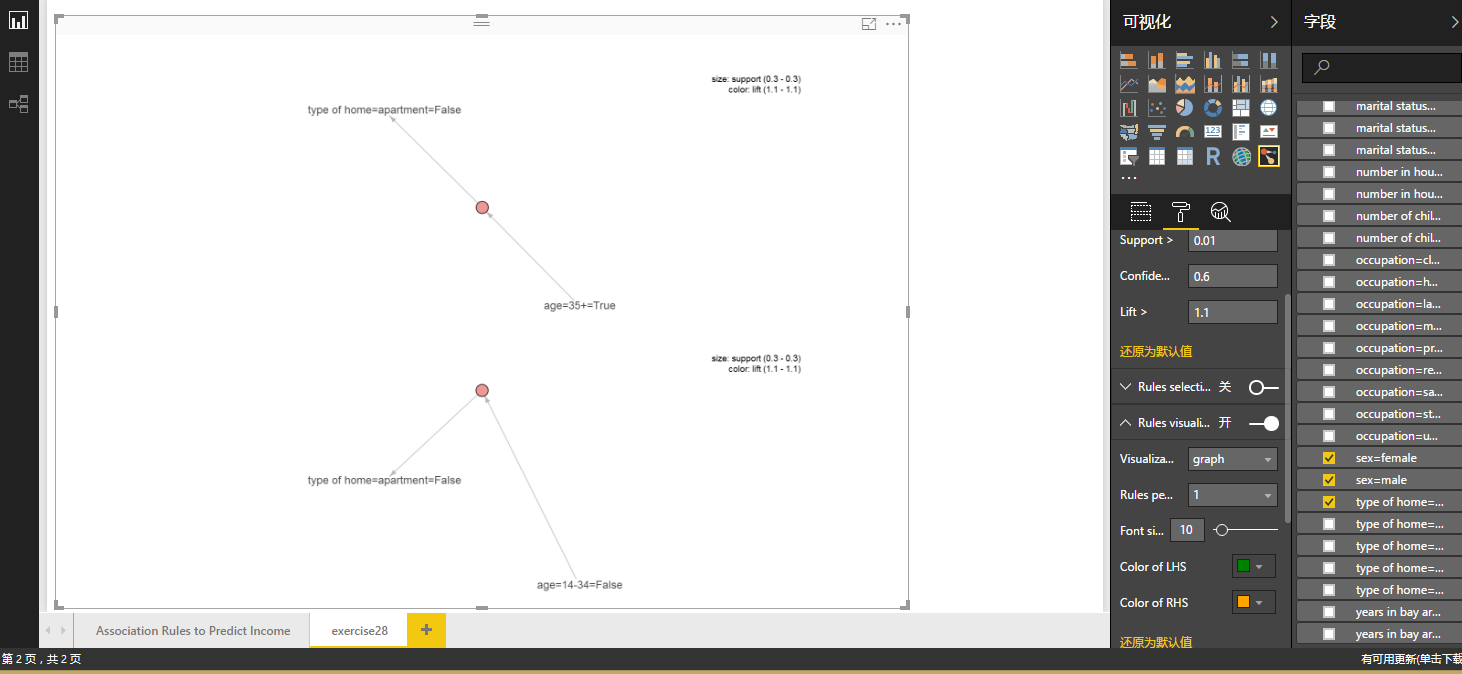
28. When the Association Rules ShowCase.pbix is opened and Advance analysis: ‘Association rules’ is stored, it shows in figure below:



I use these columns to see the association rules;



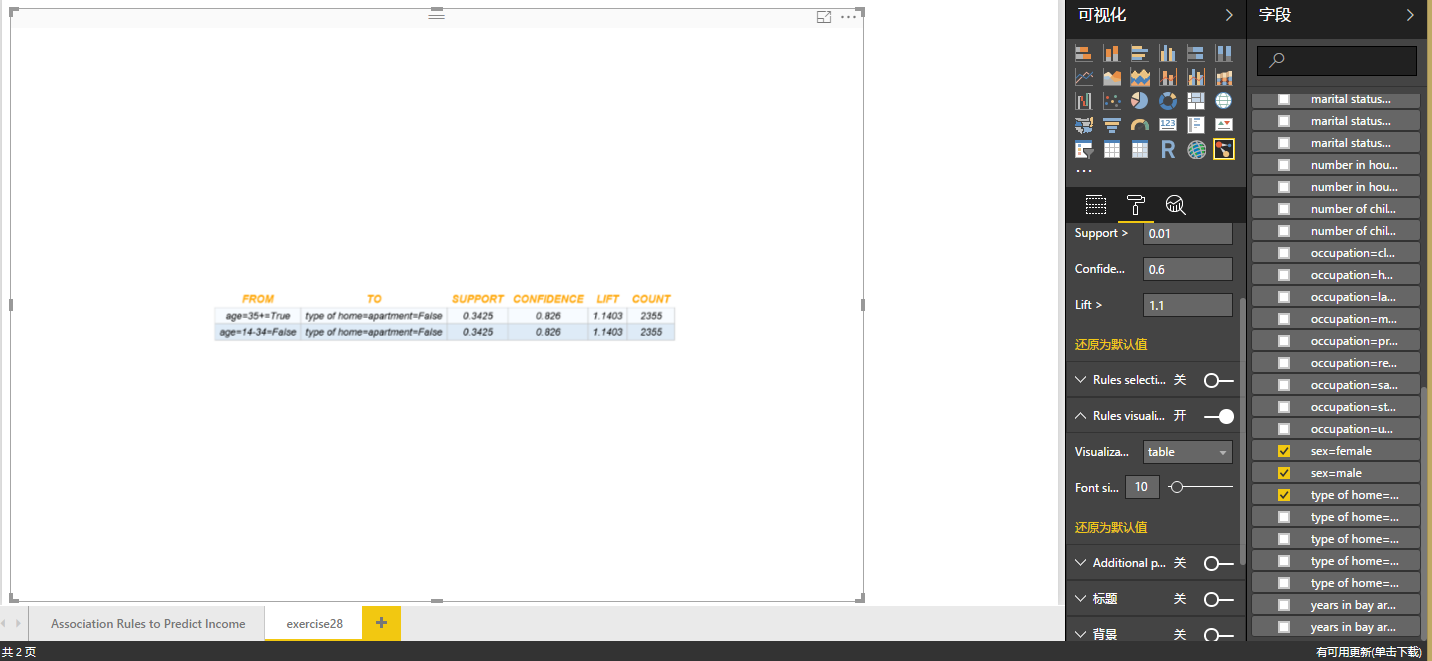
Then we get first type: Graph



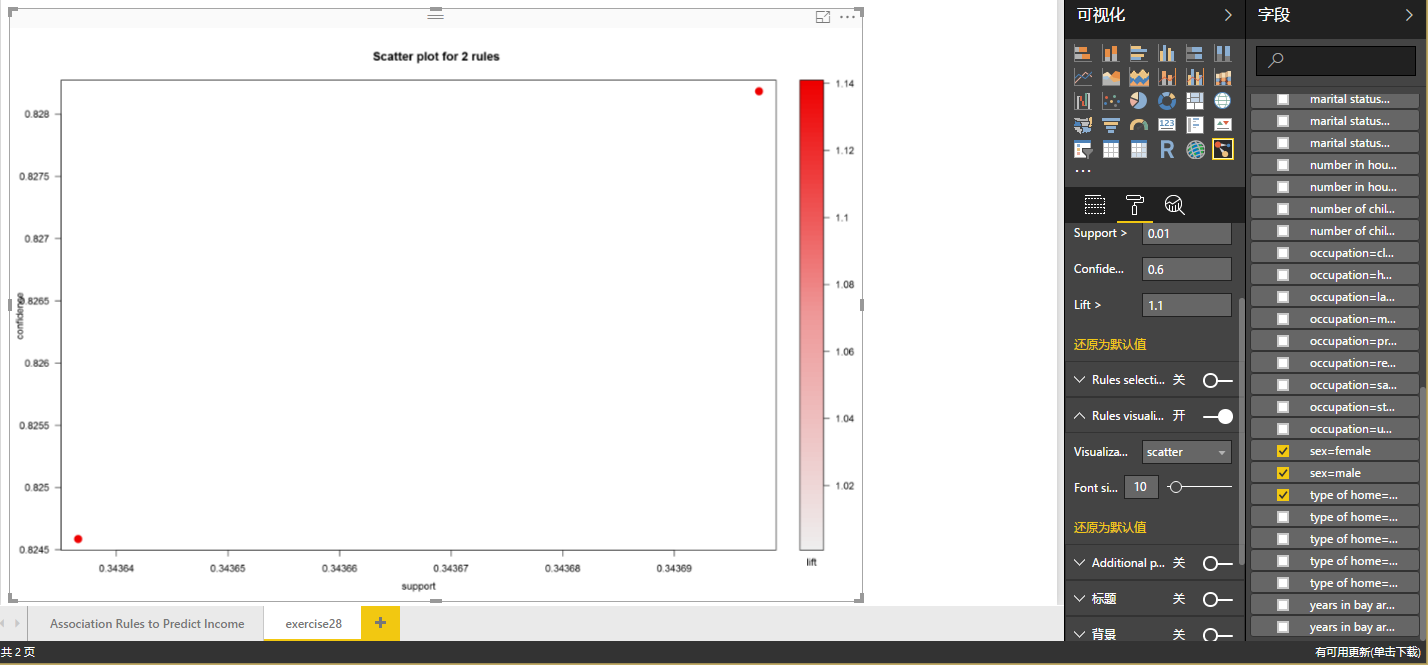
Second type: paracord



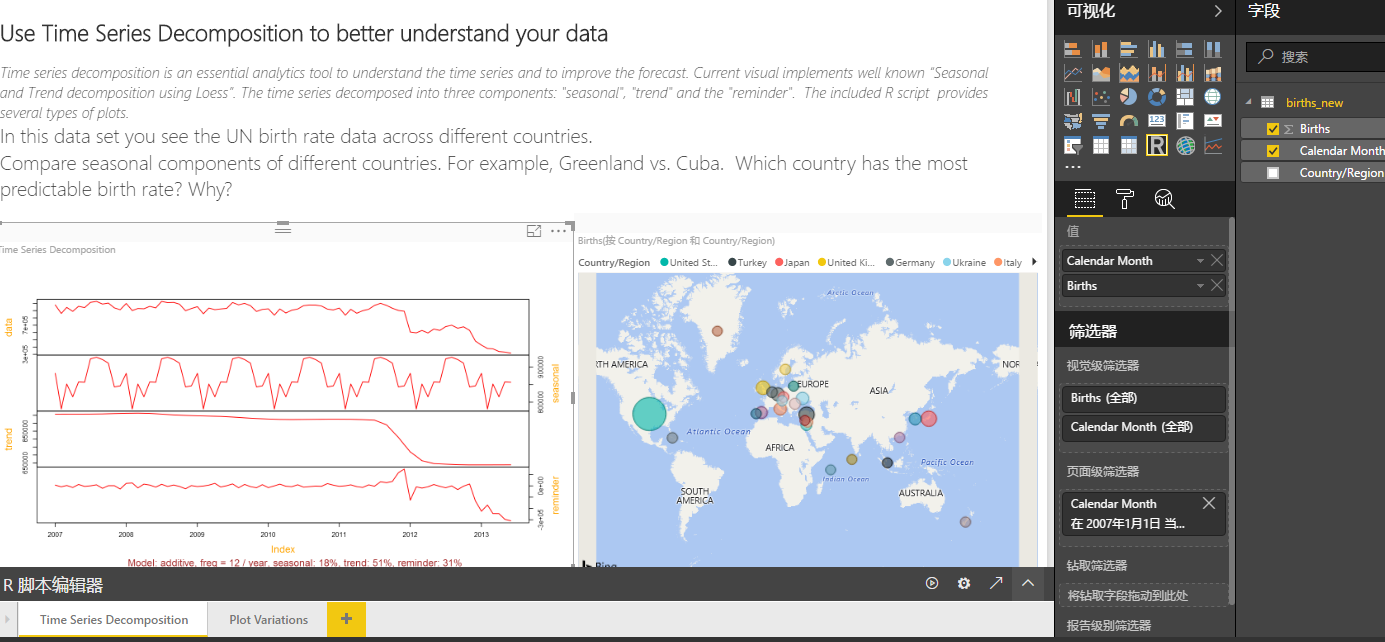
Third type: table



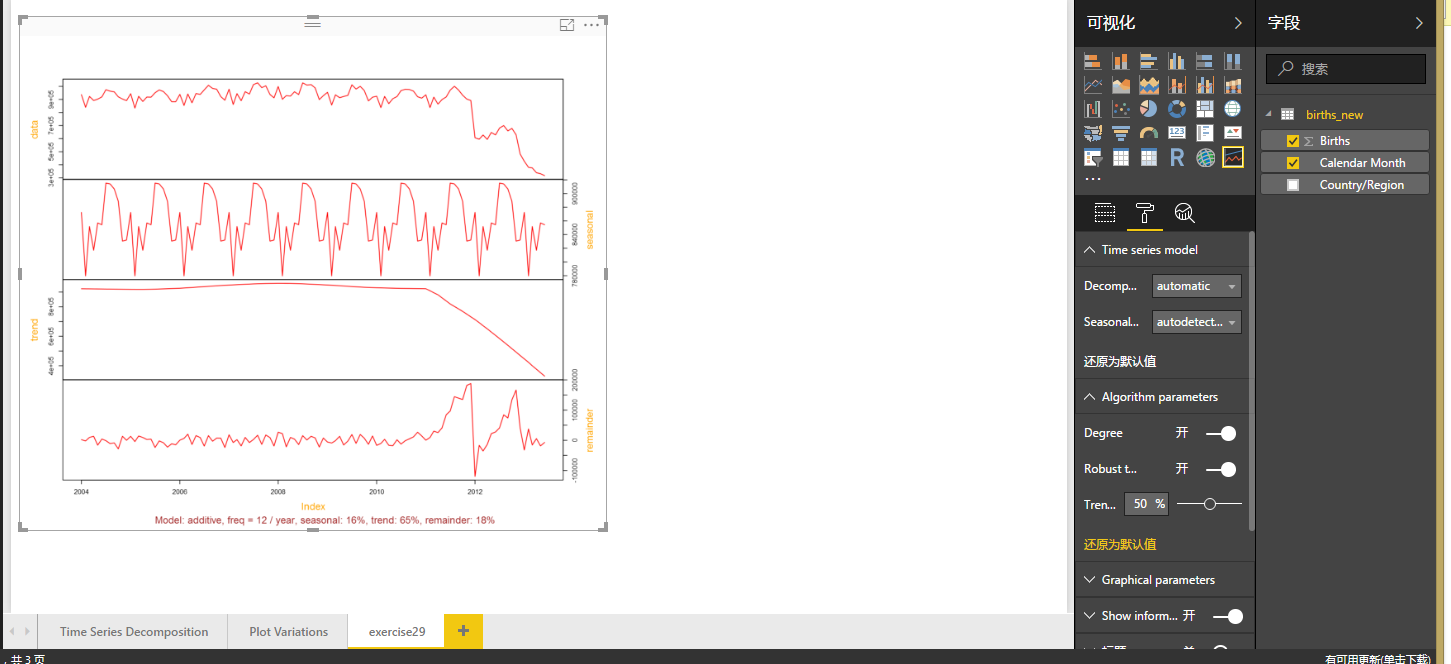
Fourth type: Scatter



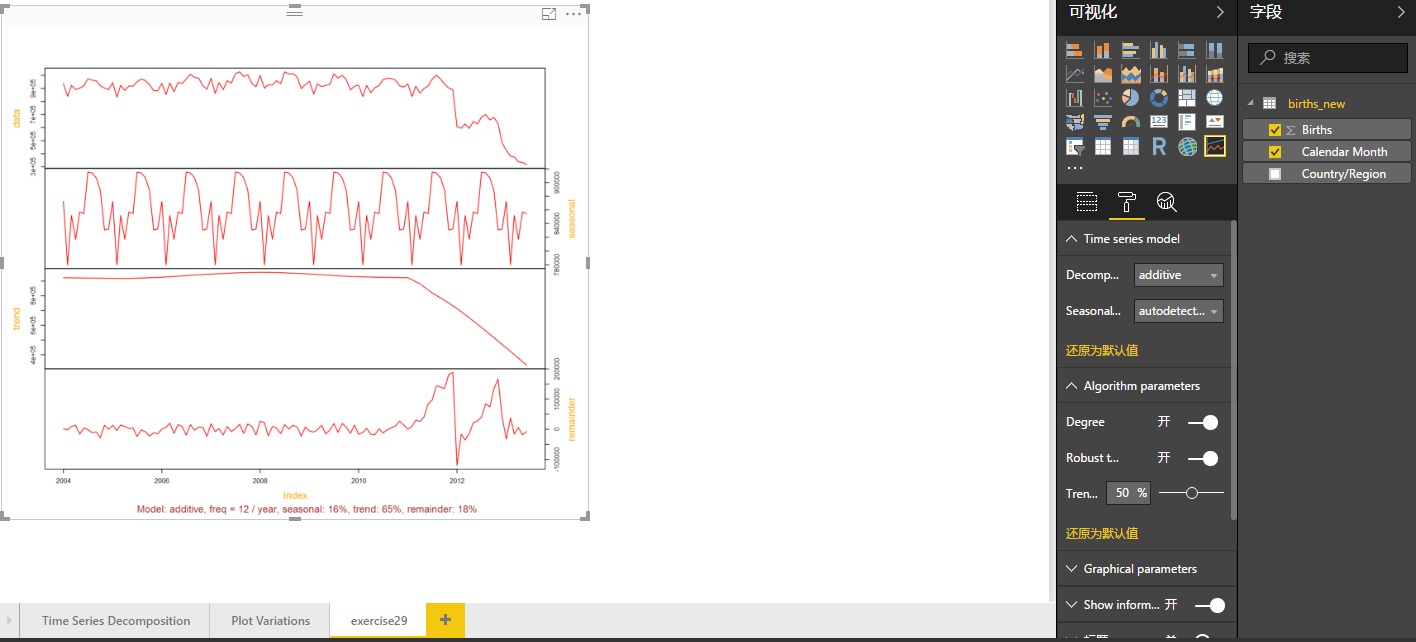
29. When the file is opened, we can see:



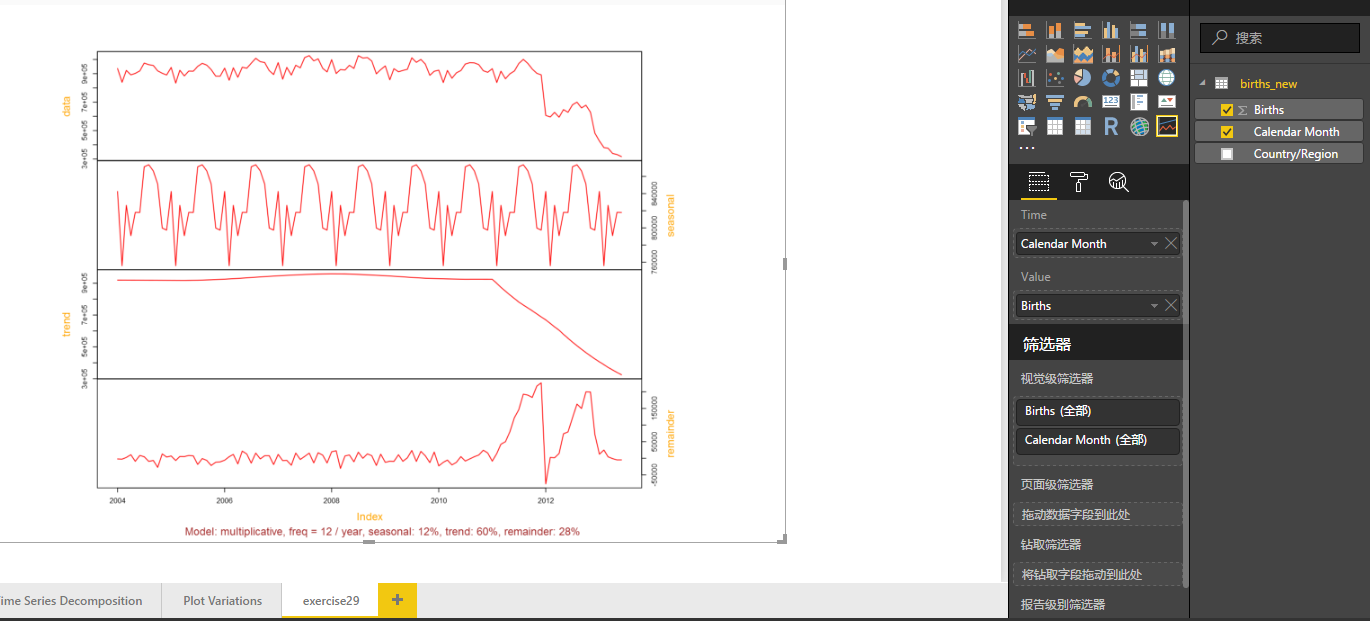


Apply Format settings Time series model and Algorithm parameters

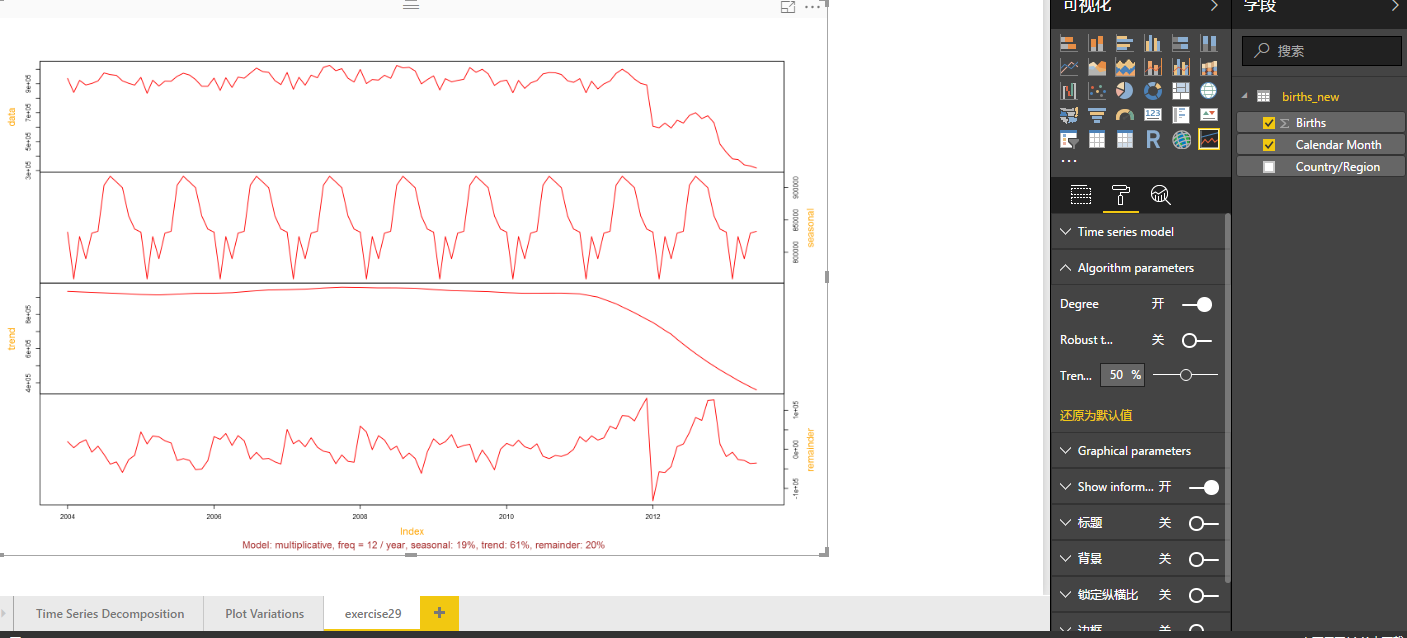
Exemplify Decomposition model additive:



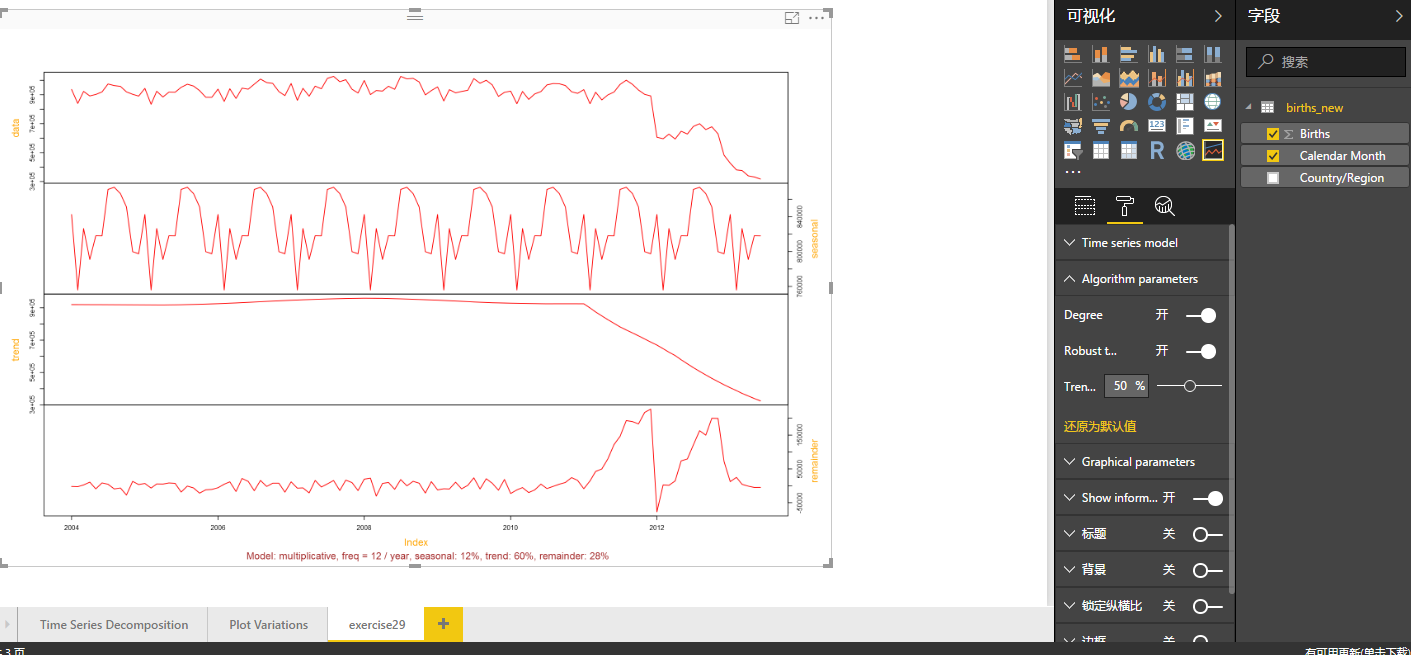
Exemplify Decomposition model multiplicative:



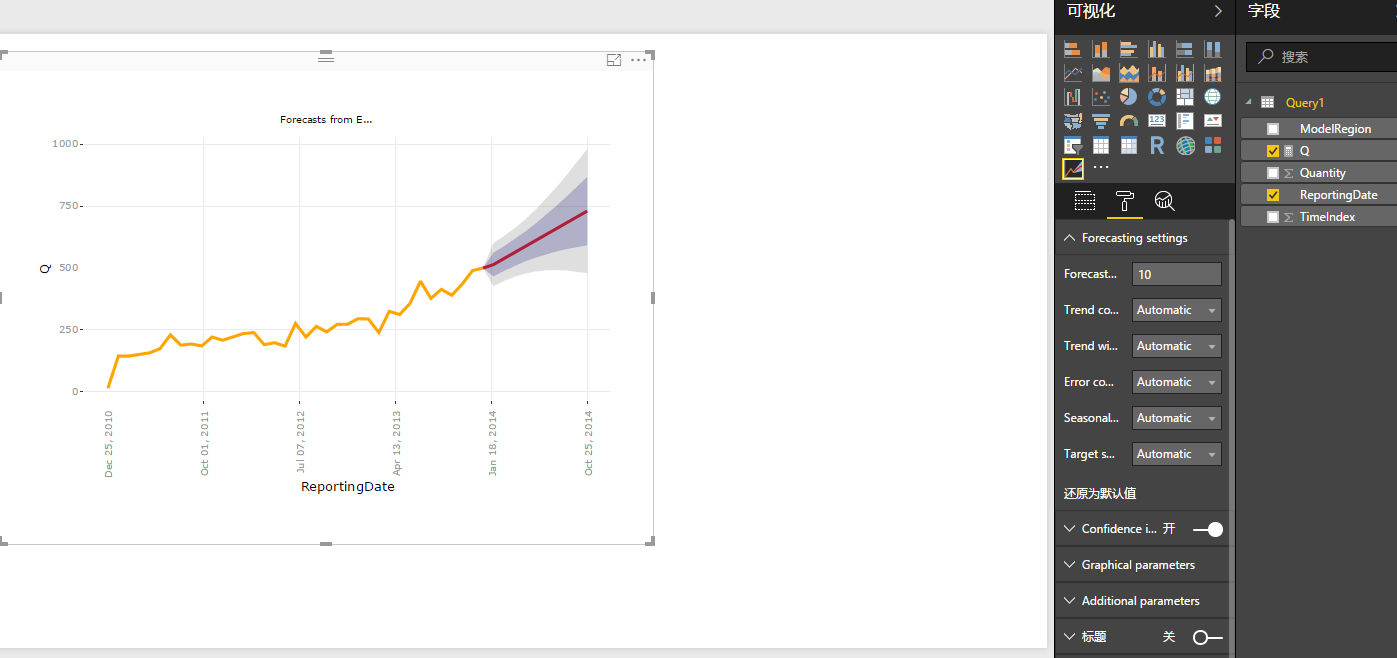
When ‘robust to outliers’ is closed, it shows like:



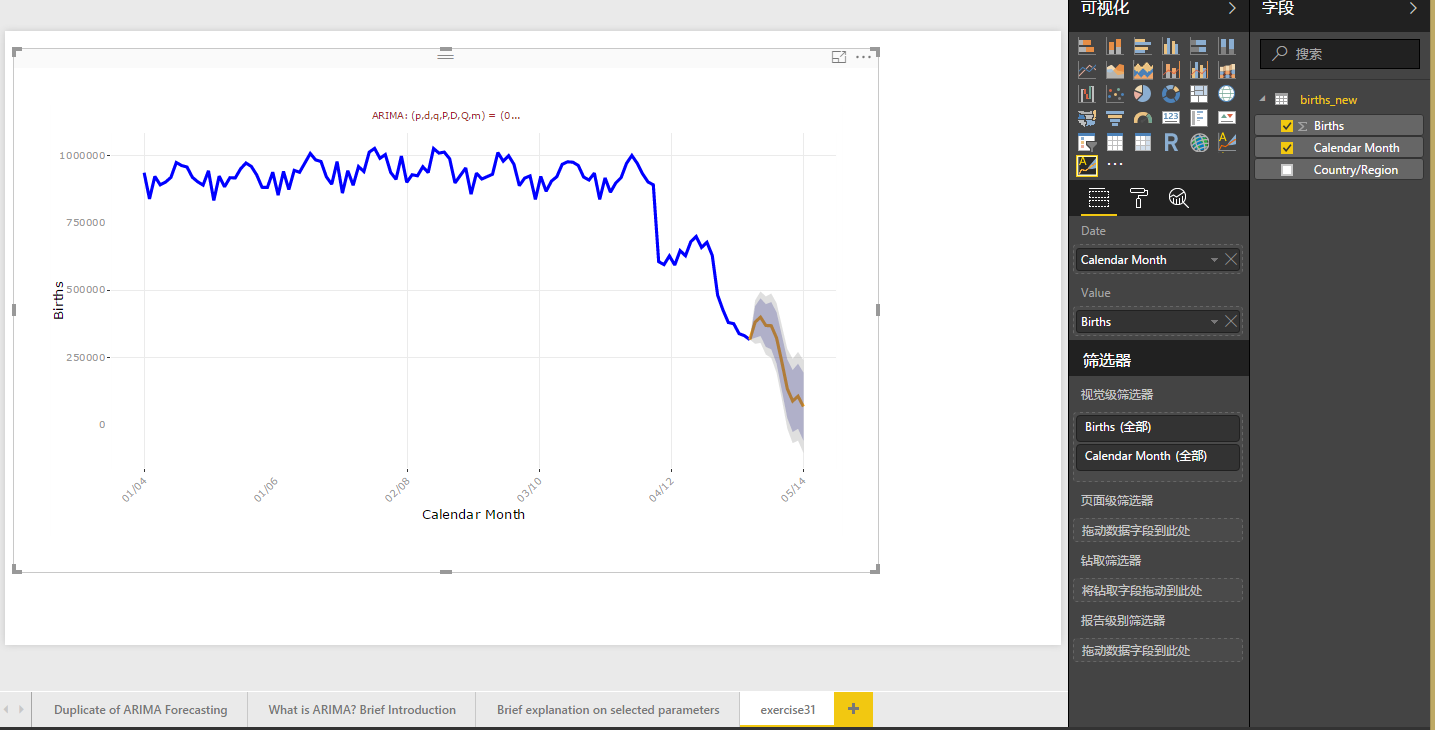
When it is opened:

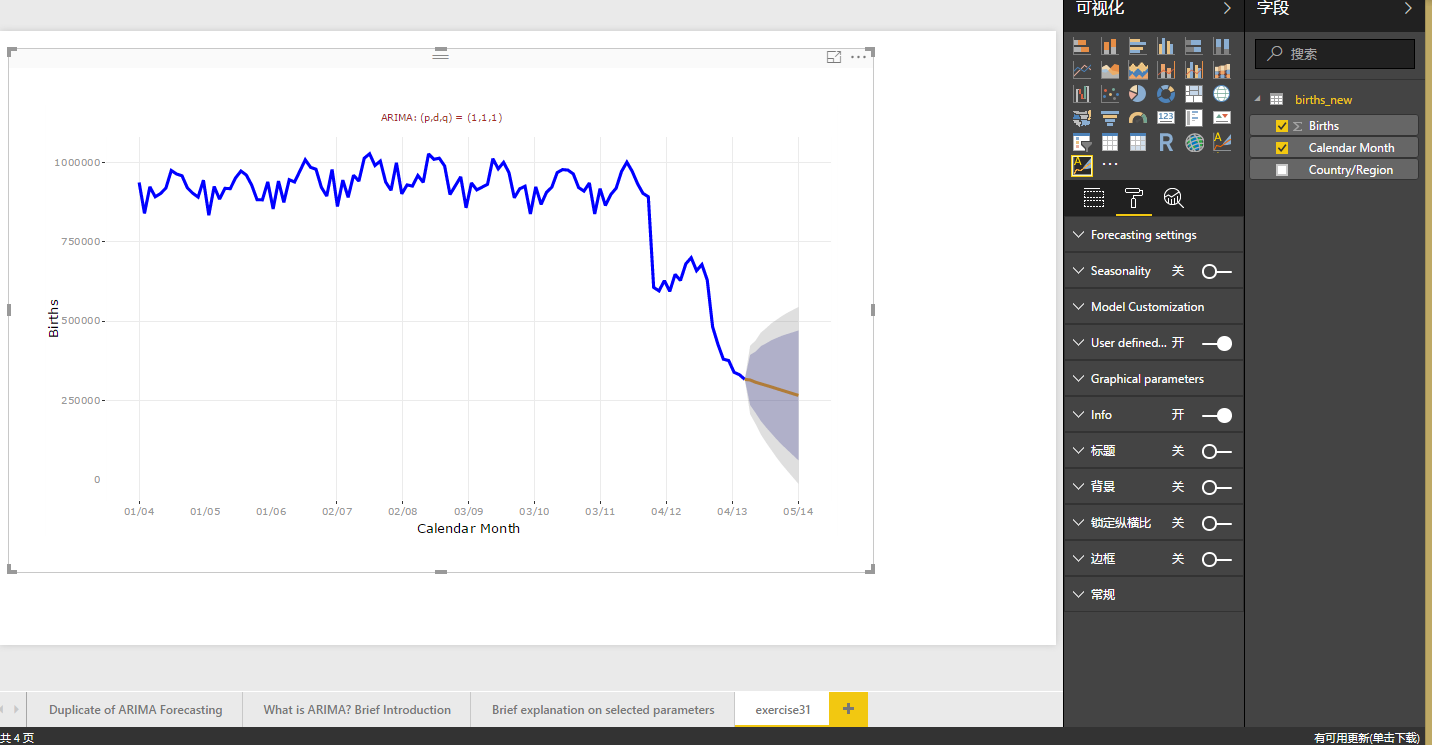


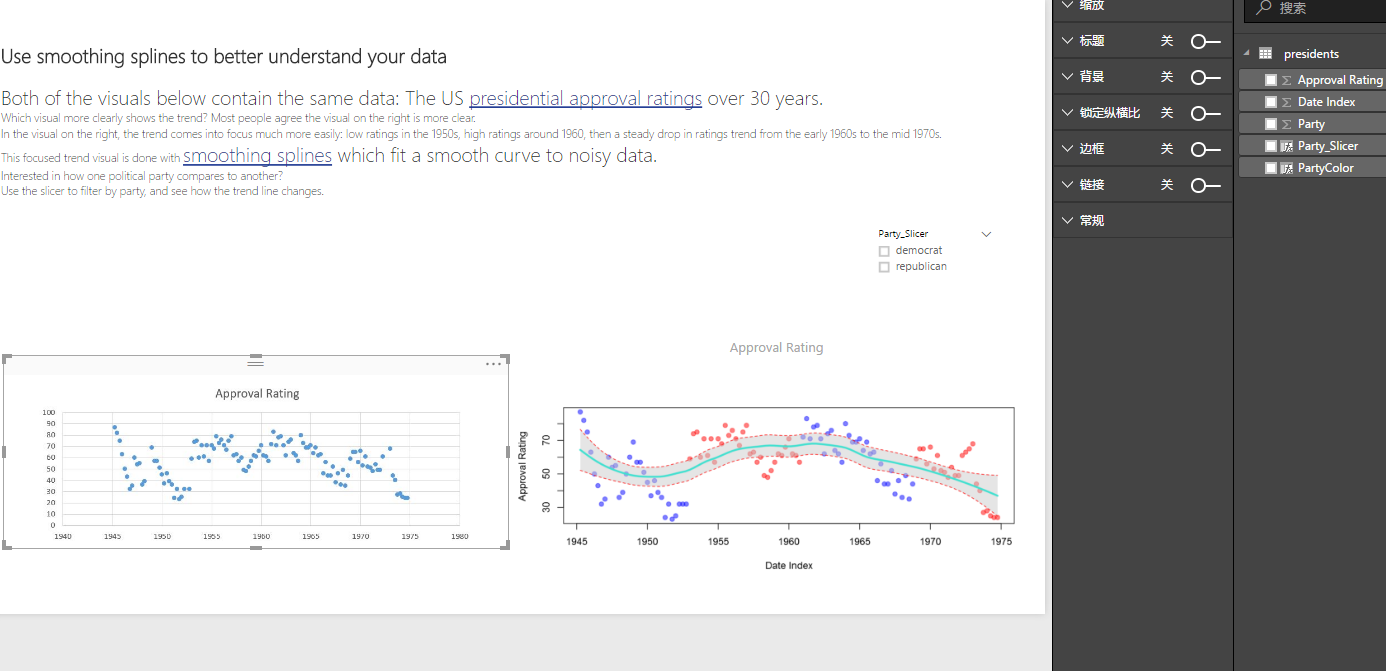
30. Make forecasts and exemplify Forecasting settings



31. Make a forecasts in new page



Apply Format settings for User defined model and set Sesonality = Off

33. 

I think the second one more clearly shows the trend, it focus much more easily: low ratings in the 1950s, high ratings around 1960, then a steady drop in ratings trend from the early 1960s to the mid 1970s.

Scatter charts are a great choice:

to show relationships between 2 (scatter) or 3 (bubble) numerical values.

To plot two groups of numbers as one series of xy coordinates.

instead of a line chart when you want to change the scale of the horizontal axis

to turn the horizontal axis into a logarithmic scale.

to display worksheet data that includes pairs or grouped sets of values. In a scatter chart, you can adjust the independent scales of the axes to reveal more information about the grouped values.

to show patterns in large sets of data, for example by showing linear or non-linear trends, clusters, and outliers.

to compare large numbers of data points without regard to time The more data that you include in a scatter chart, the better the comparisons that you can make.

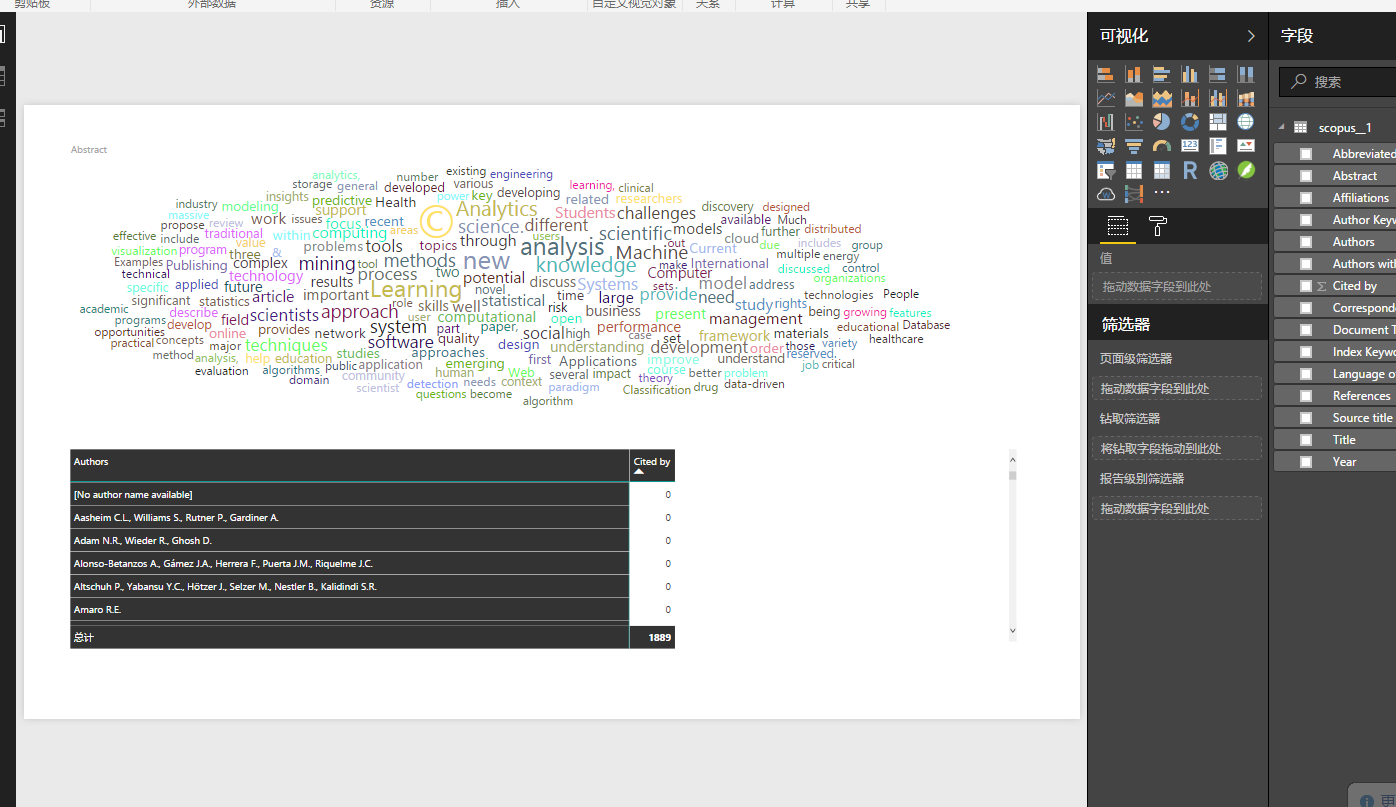
Bubble charts are a great choice:

if your data has 3 data series that each contain a set of values.

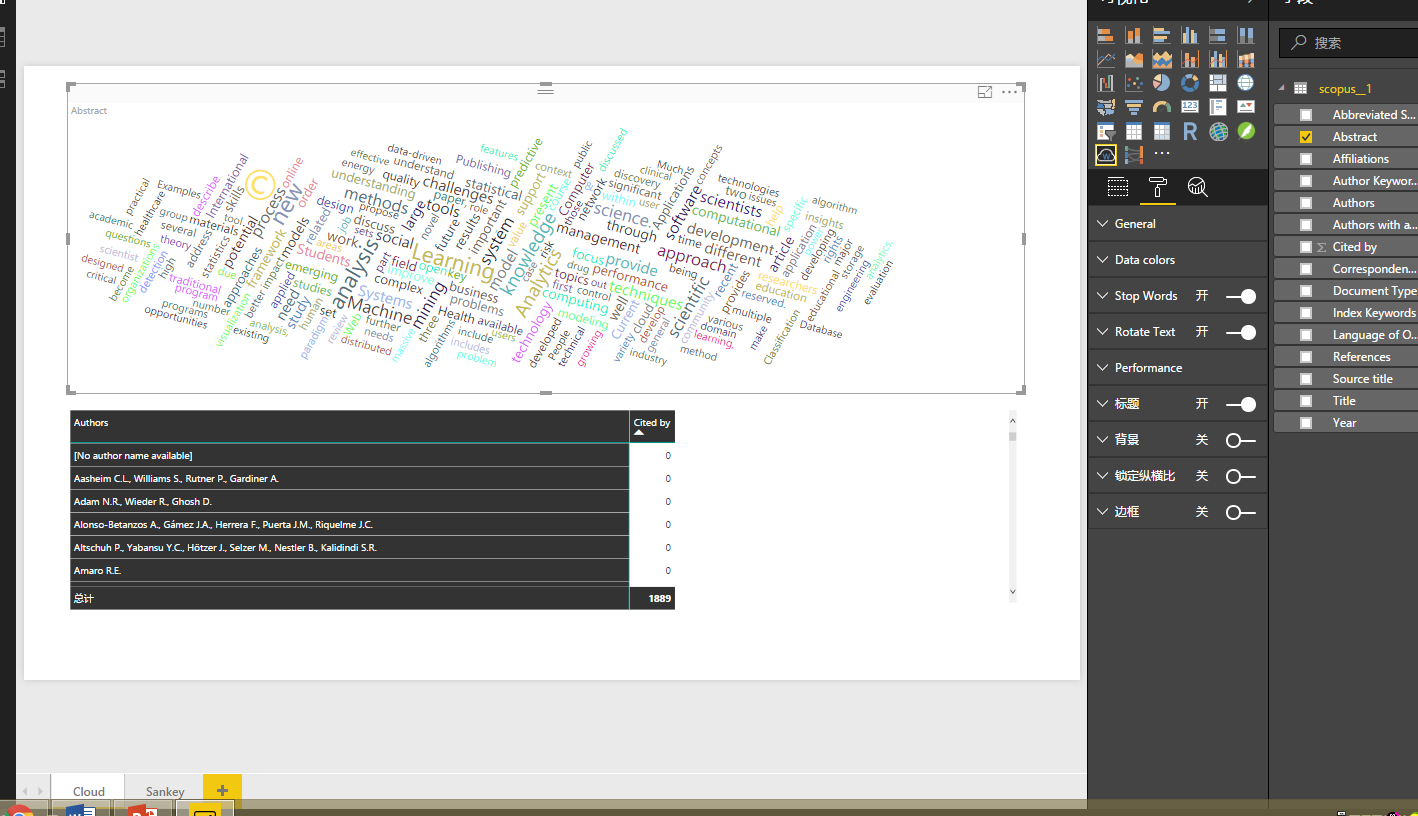
to present financial data. Different bubble sizes are useful to visually emphasize specific values.

to use with quadrants.

34. When the file is opened, we can see:



When rotate text button is opened, it shows like this:



35. :

36.

Decision Trees

Should be faster once trained (although both algorithms can train slowly depending on exact algorithm and the amount/dimensionality of the data). This is because a decision tree inherently "throws away" the input features that it doesn't find useful, whereas a neural net will use them all unless you do some feature selection as a pre-processing step.

If it is important to understand what the model is doing, the trees are very interpretable.

Only model functions which are axis-parallel splits of the data, which may not be the case.

You probably want to be sure to prune the tree to avoid over-fitting.

Neural Nets

Slower (both for training and classification), and less interpretable.

If your data arrives in a stream, you can do incremental updates with stochastic gradient descent (unlike decision trees, which use inherently batch-learning algorithms).

Can model more arbitrary functions (nonlinear interactions, etc.) and therefore might be more accurate, provided there is enough training data. But it can be prone to over-fitting as well.

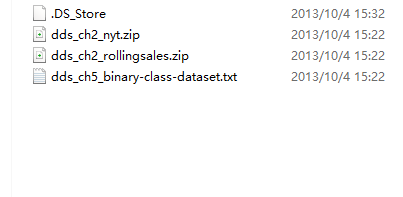
37.

38.

39. the Datasets on the book site:

<https://github.com/oreillymedia/doing_data_science>

has been choosen



40. 