Tableau Loves R: Principal Component Analysis (PCA) with Tableau R Integration

Published by Data Psycho on August 19, 2016

[Tableau, R, Principle Component Analysis, PCA]

July-07-2016:

I had a meeting with my team leader about one of our pending project. So he explained almost every thing to me. I was trying to understand all of his explanation but I barely understand fifty percent of it. The project is one of the super confidential one because of its data. where we have to work with internal official data such as employee information, projects, billing hours, video conferences and travel information. I will not be able to share the original data and related information in this post. But do not worry I will explain exactly every thing what I have done through a dummy data. Also I will share the dummy data so that you can apply all the methods step by step what I did.

So my team always works with internal official data, not directly with the customers and our main customers are the Managers, CEO, budget team and other higher ranked personals who wants to see some analytics about their team or budget to take some decision. First let me explain the project in short. The management of my company want to emphasize on video conference between the offices all over the globe, which will reduce the travel of the consultant (what they are thinking). That's how consultant can save their time (though consultant wants to travel) and company will save the money. So it's all about money again. For me every technology was new what we are using in this project mainly tableau and Alteryx.

[DATA]

July-08-2016:

I start to work on the project. I had five different data sources, some of them are based on Oracle SQL Server and some are Microsoft SQL Server. I will not explain about it in this blog which can make it boring. But this data integration part help me to realize the truth that, "You will spend 80% of you time on data wrangling". Because It was almost 10 days and I was struggling to to make the whole ETL process in Alteryx, which should provides an output of Tableau readable (.TDE) file. Which will be used (.TDE) in Tableau Desktop to make some visual analytics. In the next of my section I will just explain a small part which is about PCA. I will also tell about the mistakes I did through my whole work, so that you can learn from my mistakes.

July-18-2016:

In the second meeting I proposed my leader to include PCA in our visual analytics part. I just proposed but still then I did not have any idea how to do it in Tableau. He was very happy to include it. So I take it as a challenge because the data was perfect for PCA and not so many people were using the tableau R interrogation. I had a good experience in R but I did not have any idea about Tableau R integration. But These resources helps me a lot in the primary stage.

Tableau and R integration in Four Steps, Tableau and R Overview.

After following these resources I was successfully able to establish the Tableau R Integration. I personally call it TRI which will be use in the next of my writing.

July-26-2016:

I used PCA a long time ago when I was in university. So I see some YouTube video which helps me to understand some basic idea. You can watch it in here. Then I start searching some resources about TRI from where I can learn how to do in step by step. But as usual there is no 100 percent complete resource. Other wise I did not write my first blog ever. The best help I found from this blog. I spent almost twenty hour on all of the resources.

August-01-2016:

I load the data in to R and start exploring everything. The proportion of variance plot I get from R was satisfactory. The first and second component covers almost all variability. But I have just three variables so it is very usual. Then I start to follow this blog step by step and download the .twbx file to have a look on it. Then I discover some of the worksheet of this file is not working and throwing some error. So I start to find some solution. I almost forget about my own project and keep searching the solutions which I did not find still now. That's how I spent another week of 20 hours without having any improvement.

August-08-2016:

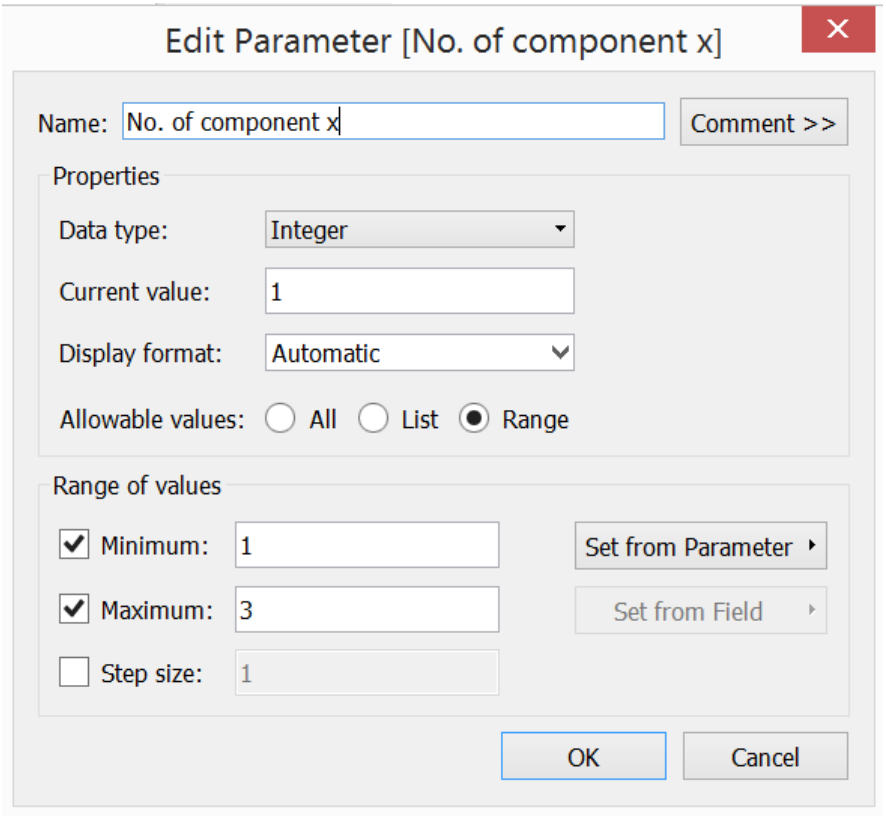
I start working my own data. I load the data in to tableau from a excel file. I start to follow the instruction from the blog and apply that on my data. First I start to create the calculated field which is used to store the principle component in tableau. The following image show the script for calculating the first component:

component x

Results are computed along Table (Across).

```
SCRIPT_REAL('
ncomp <- max(.arg4)
df <- data.frame(.arg1,.arg2,.arg3)
pc <- princomp(df, cor = TRUE)
as.numeric(pc$scores[,ncomp])
',
SUM([Total Bl Hours]),SUM([Total Fl Milage]),SUM([Total Vc Minute]),[No. of component x])
```

We can see tableau use SCRIPT_REAL() to contain the main R code and taking the aggregate value of main fields as argument. There are lot of different ways to write the code. I am using the basic one. But my script through an error for the last argument. Still then I did not know that, what kind of argument is that (As I copied and pasted the script from some source like some of you). After some hours of research, I am able to familiar with a feature of tableau which is called Parameter Control. Then I create two required parameter for my two axis like the following image:

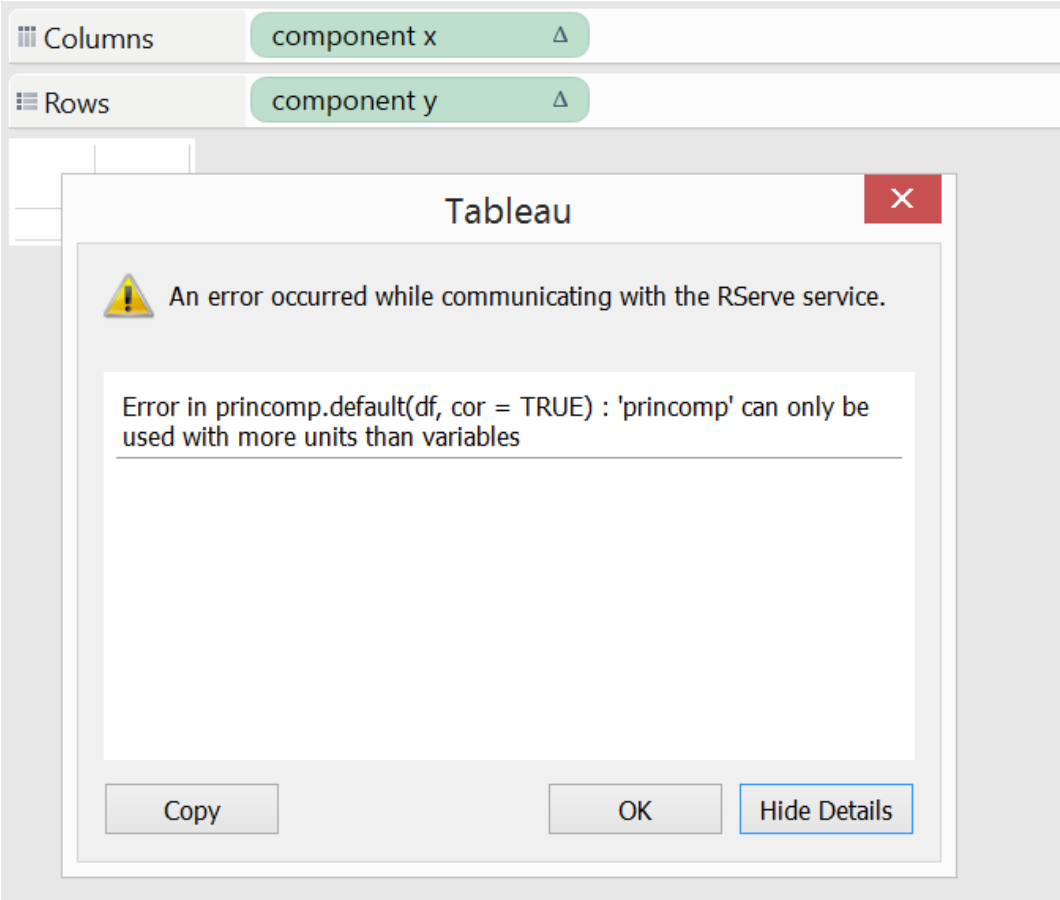


As I will have maximum 3 principle component so I set the parameter for 3. Then I create another parameter for Y axis component using the same setting but with a new name. Then when I go to the calculated field it does not show any error. The same script for calculating component of Y axis is as follows (I know you are lazy):

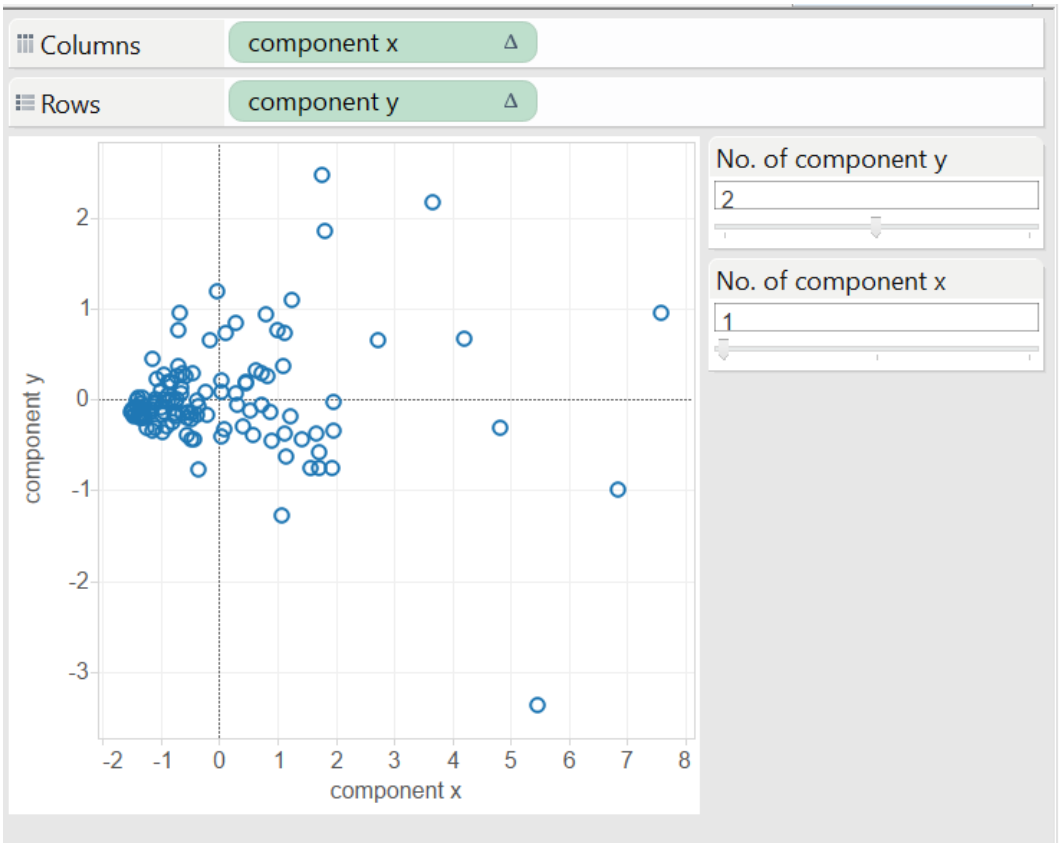
```
SCRIPT_REAL('
ncomp &amp;lt;- max(.arg4)
df &amp;lt;- data.frame(.arg1,.arg2,.arg3)
pc &amp;lt;- princomp(df, cor = TRUE)
as.numeric(pc$scores[,ncomp])
',
SUM([Total Bl Hours]),SUM([Total Fl Milage]),SUM([Total
```

August-09-2016:

Most probably one of the happiest person in the office. He is going to finished this PCA (here I am rock you like a hurricane !!! I am a creep I am weirdo..ooo.... !!!). He open his laptop **first run the R server** package (do not forget to do it) then turn on the Tableau and put the calculated field in the row and column bar in table. What did he see?:



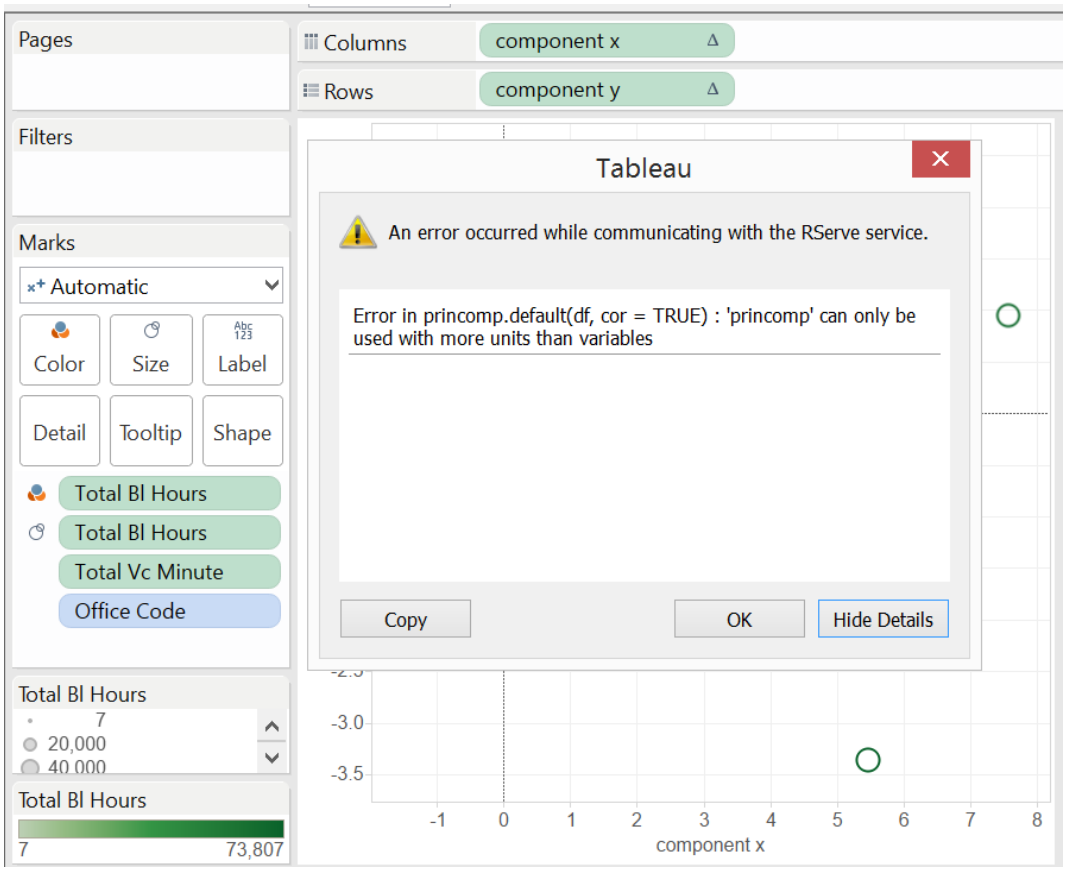
So the dreams became nightmare for me. I start to figure out my mistake. I search all available resources but I could not find any solution (or I was stupid and retarded). Next day at the end of my office hour I discover from a stack overflow post that, Tableau by default use aggregation on fields. which creates “variable more than rows (column count> Row count)” for R calculation in my case. The concept is, variable is greater than the observation. So I turn off the default aggregation and finally able to plot the principles components. Then I also activate the parameter control in the right pan and plot the first two component.



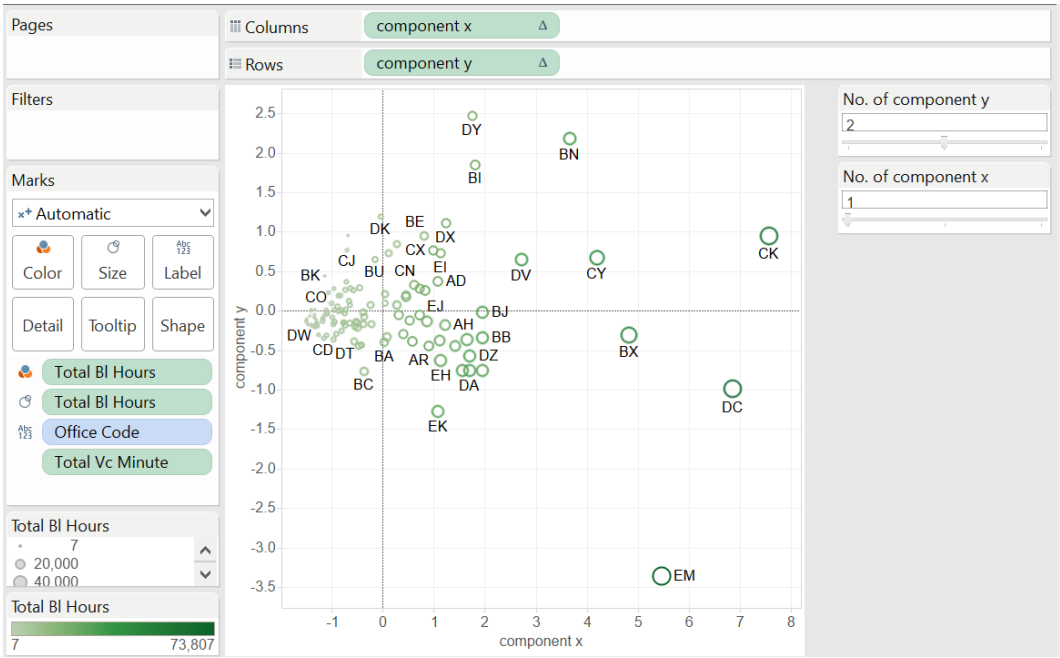
So finally I finish the PCA in Tableau.

August-15-2016:

I add my original fiends Total Flight Hours, Total Bill and Total Video Conference in to Size, Color and Detail filter. Then I put the Office Code in Label and I get this:



Then I was thinking myself as the most stupid person in the world. I spent more one and half day to figure out what is the problem. Again I have been Introduced with a new feature of Tableau. By default it use Measure option on dimensions when you drag them to some function. So when I drag Office Code to Label, Tableau send it to R for calculation of PCA. So R through an error which I got before. I change the property of Office Code from Marks pane Measure to Attribute, fortunately it solved all the problem and the nightmare become festival for me (Eid, Puja, Christmas, Purnima, Sabbat, New Year) !!!



Now I am able to see the relative position of the Offices in terms of first and second principle component. That's how I am able finish this PCA with TRI successfully on probably 17th of August.

"I spent more that 50 hours to finish this PCA with other assigned projects. Some time I feel stupid, some time hopeless some time tired and some time successful. But I never gave up."

Published in Tableau

Tableau Loves R

No Older Posts
Return to Blog

Next Post
Alteryx Good, Bad & Ugly: Introduction & Lets Create a Template