Lab 8

TabPy

Spring 2019 DS-1001

TabPy

- https://github.com/tableau/TabPy
- Download the latest version from: https://github.com/tableau/TabPy/releases
- Unzip

Starting up TabPy

Mac

1. Open a terminal.
2. Navigate to the folder in which you downloaded your source code.
    - This folder should contain the file: `startup.sh`
3. Run the following command from the terminal:

```
./startup.sh
```

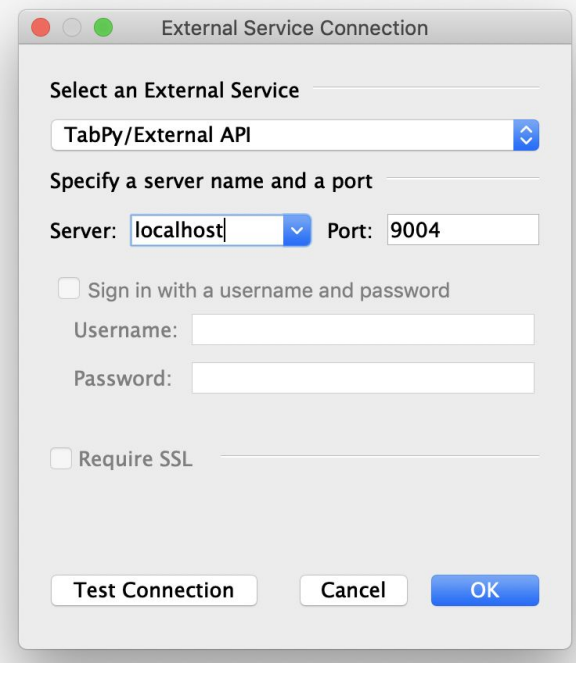
Windows

1. Open a command prompt.
2. Navigate to the folder in which you downloaded your source code.
    - This folder should contain the file: `startup.cmd`
3. Run the following command from the command prompt:

```
startup.cmd
```

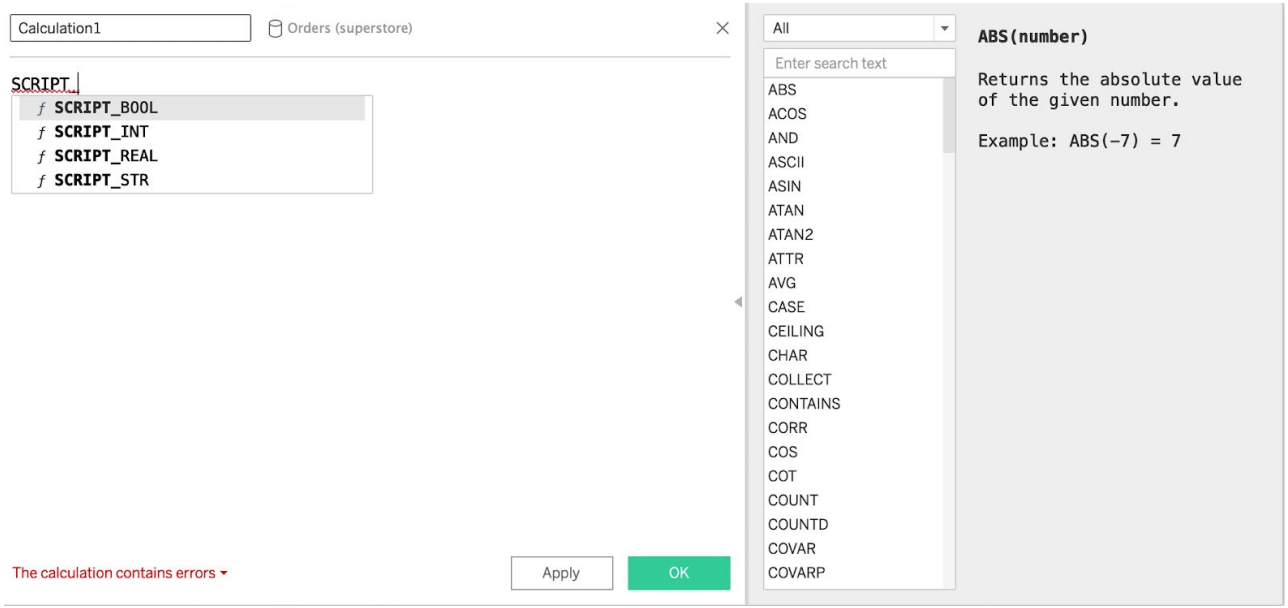
https://github.com/tableau/TabPy/blob/master/docs/server-startup.md

Connect Tableau Desktop to TabPy server

1. Open Tableau Desktop
2. Help -> Settings and Performance->Manage External Service Connection
3. 

Create Calculated Fields with Python code

The calculated field accept returns as **a single number** or **a list** of values, which can be Boolean, Integers, Real numbers or String. Specify the output data types by SCRIPT_XXXX



Basic Structure of Calculated Field with Python

Tableau TabPy:

```
SCRIPT_BOOL("

profitable = []

for x in _arg1:

        profitable.append(x>0)

return profitable",

SUM([Profit]))
```

Must be an aggregated measure.

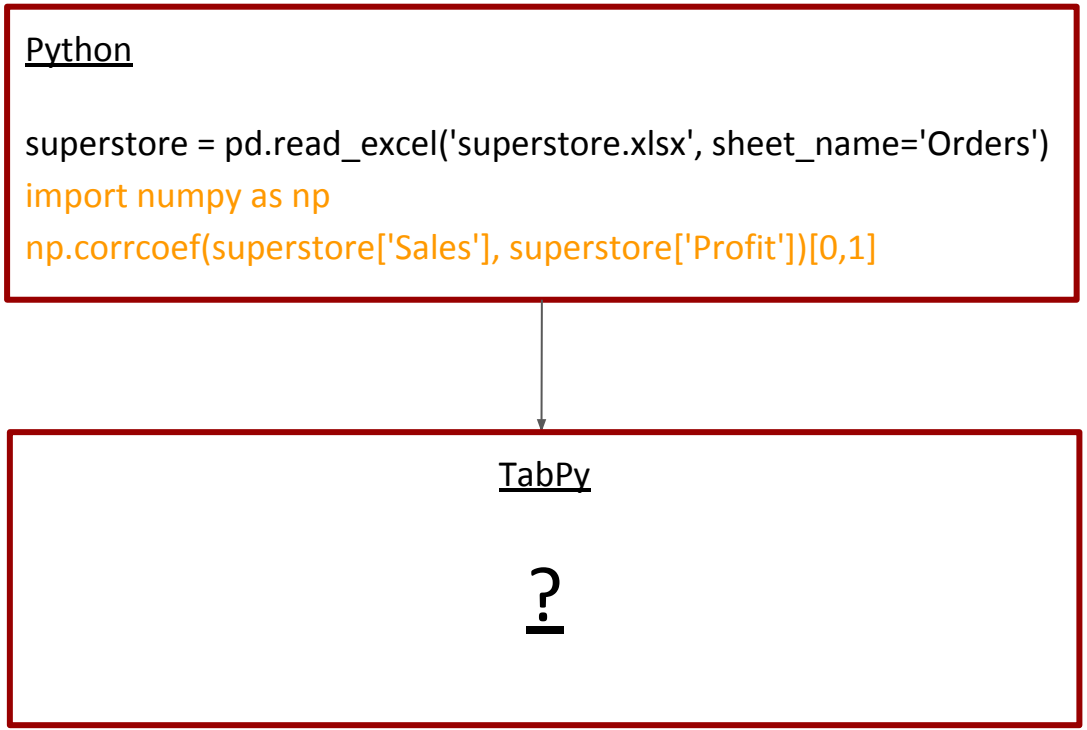
Python

```
profitable = []

For x in profit:

        profitable.append(x>0)
```

Compute Correlation Coefficients



Plot


Data
Analytics
airline
Orders (superstore)
seedsdataset
seedsdataset (2)
Dimensions
City
Customer ID1
Customer Name
Customer Segment
Order Date
Order ID
Order Priority
Postal Code
Product Category
Product Container
Product Name
Product Sub-Category1
Region1
Row ID
Ship Date
Ship Mode
State or Province
Measures
correlation coefficient
Discount
Product Base Margin
Profit
Profitable?
Quantity ordered new
Sales
Shipping Cost
Unit Price
Parameters
selected_pca_component
Pages
Filters
Marks
Circle
Color
Size
Label
Detail
Tooltip
Customer Name
correlation c..
Columns
Product Category
SUM(Profit)
Rows
Region1
SUM(Sales)
Sheet 11
Product Category
Region1
Furniture
Office Supplies
Technology
Central
East
South
West
Sales
100K
50K
0K
-10K
0K
10K
20K
Profit
AGG(correlation coeffic...
-0.1507
0.7392

PCA with TabPy

Things to keep in mind:

1. TabPy takes into ‘aggregate measure’ as input (eg. sum, average, …), hence, we need to **create an unique index** for every row and make the calculated fields calculate along the unique index (Edit Table Calculation).
2. TabPy Calculated Fields output only 1 list, hence for each principal component, we need to create a different calculated fields.
3. Create Parameter to point out which principal component we are looking at.

PCA in Python and TabPy

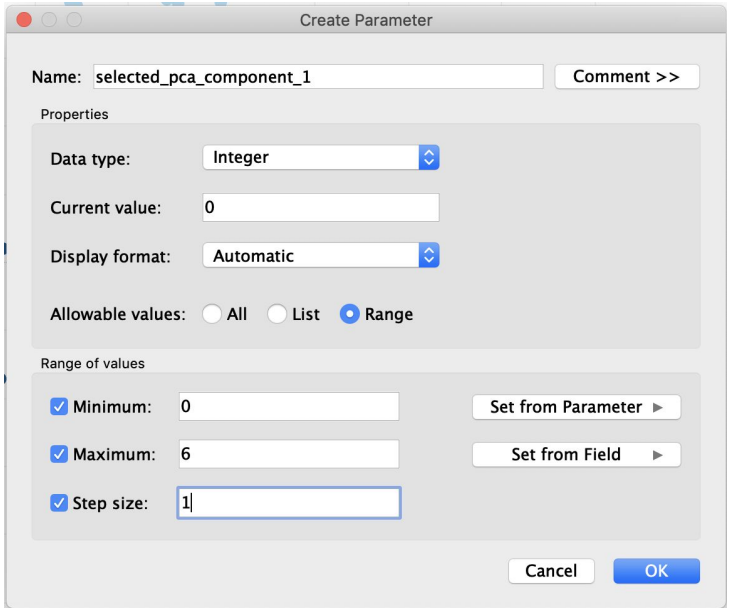
```
SCRIPT_REAL("
import pandas as pd; from sklearn.decomposition import PCA; from sklearn.preprocessing import StandardScaler

// df = pd.read_csv('seedsdataset.csv'); df = df.drop('wheat_variety', axis = 1)
df = pd.DataFrame({'ar': _arg1, 'perm': _arg2, 'compact': _arg3, 'kernel_leng': _arg4,'kernel_width': _arg5, 'asym':
_arg6, 'kernel_groove_leng': _arg7})

n_comp = len(df.columns)
scaler = StandardScaler()
scaled_df = scaler.fit_transform(df)
pca = PCA(n_components=n_comp)
pca_comp = pca.fit_transform(scaled_df)
return list(pca_comp[:, _arg8[0]])",

SUM([Area]), SUM([Perimeter]), SUM([Compactness]), SUM([Length Of Kernel]), SUM([Width Of Kernel]),
SUM([Asymmetry Coefficient]),SUM([Length Of Kernel Groove]),[selected_pca_component_1])
```

Create parameter [selected_pca_component_1]

1. Right click at Dimensions -> create parameter
2. 

3. Click on the parameter -> Show parameter control

Plot

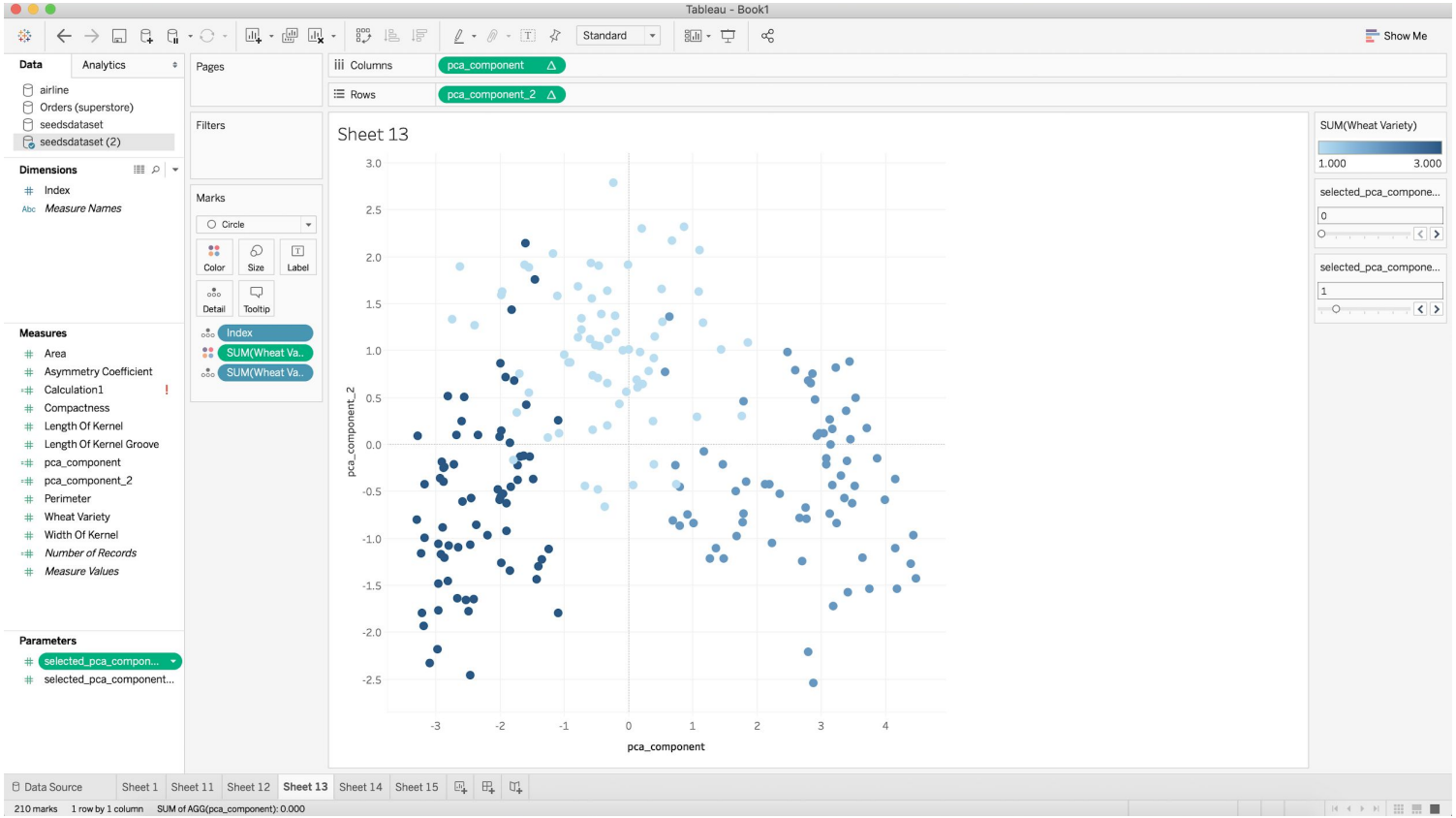


Tableau - Book1
Standard
Show Me
Data
Analytics
airline
Orders (superstore)
seedsdataset
seedsdataset (2)
Dimensions
Index
Measure Names
Measures
Area
Asymmetry Coefficient
Calculation1
Compactness
Length Of Kernel
Length Of Kernel Groove
pca_component
pca_component_2
Perimeter
Wheat Variety
Width Of Kernel
Number of Records
Measure Values
Parameters
selected_pca_compon...
selected_pca_component...
Pages
Filters
Marks
Circle
Color
Size
Label
Detail
Tooltip
Index
SUM(Wheat Va..
SUM(Wheat Va..
Columns
pca_component
Rows
pca_component_2
Sheet 13
pca_component_2
pca_component
SUM(Wheat Variety)
1.000
3.000
selected_pca_compone...
0
selected_pca_compone...
1
Data Source
Sheet 1
Sheet 11
Sheet 12
Sheet 13
Sheet 14
Sheet 15
210 marks
1 row by 1 column
SUM of AGG(pca_component): 0.000

Sentiment Analysis with TabPy - billboard.csv

1. First, download necessary nltk package
   (IN PYTHON): **import nltk; nltk.download('vader_lexicon')**
   (In Command Line): **pip install twython**
2. https://gist.github.com/andrea-w-wang/b2b34e6d3a8d1fe513ecde6dc303d06b

Reference and Additional Resources

- Reference
  - http://alexloth.com/2016/11/06/tabpy-tutorial-integrating-python-tableau-advanced-analytics/
  - https://tc18.tableau.com/sites/default/files/session/assets/18BI-009_Building_Data_Science_Applications_in_Tableau.pdf
  - VADER:
    Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
  - http://www.nltk.org/howto/sentiment.html
  - https://boraberan.wordpress.com/2013/12/24/sentiment-analysis-in-tableau-with-r/
  - https://gist.github.com/databrit
  - Billboard dataset; seedsdataset; superstore dataset
- Additional Exercise/Examples of TabPy
  - Forecasting with TabPy, Machine Learning in TabPy