Machine Learning Using Splunk And R

Marianne Faro

Daniel Koops

Gijs Wobben

Itility

Disclaimer

During the course of this presentation, we may make forward looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not, be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Agenda

- About Itility
- Machine Learning Use Case
- Architecture And Challenges
- The New App
- Demo

About Itility

Our business

- IT consultancy
  - Data Science
  - IT infrastructure
  - Software development
  - Datacenter automation
  - IoT
  - …

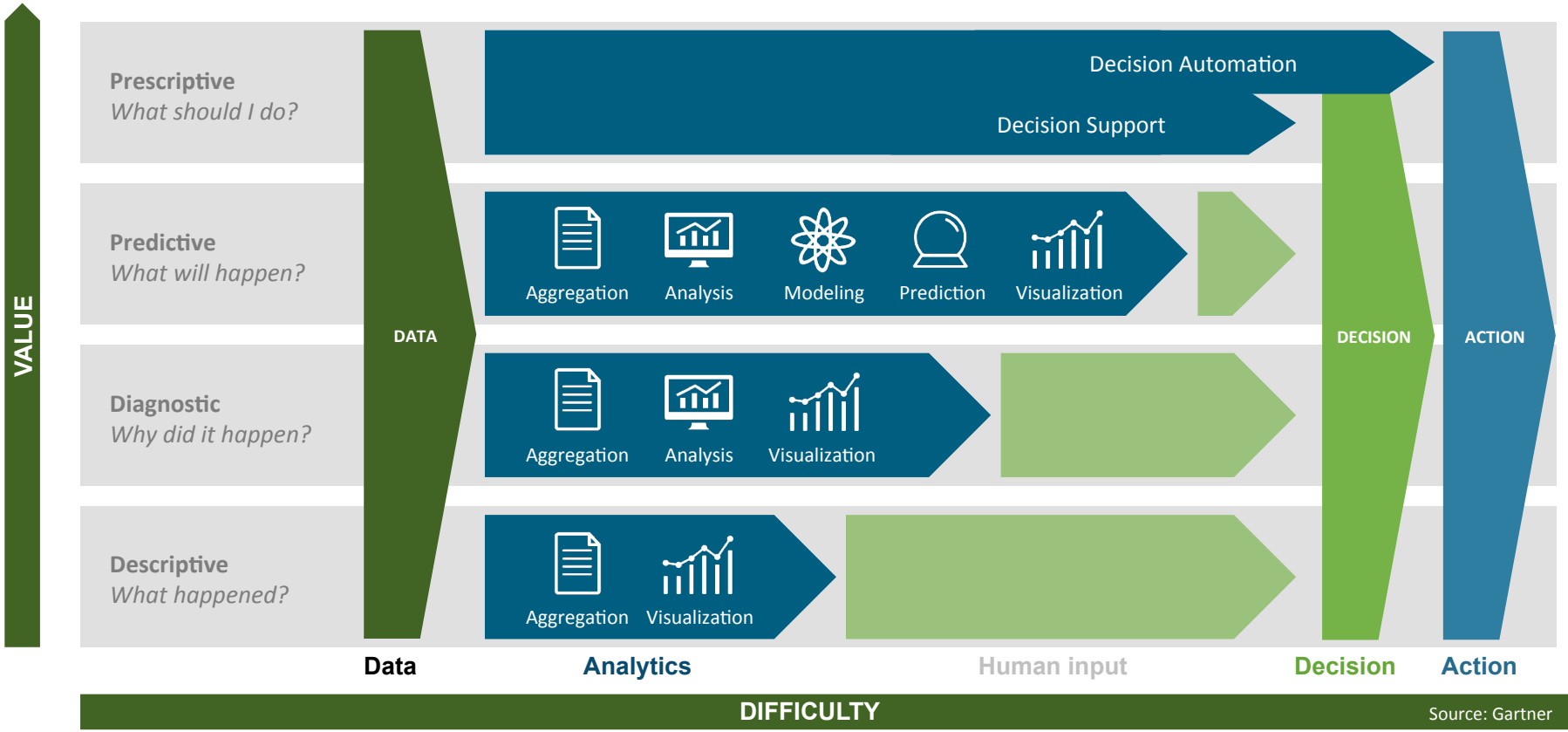
Our Customers

Data Science Goals

- We help our customers gain visibility into their data
- We use data to diagnose incidents and find root causes
- We predict and forecast to anticipate upcoming issues and problems
- We automate decisions to act on upcoming issues and problems

- … and we do so by using everything between basic statistics and the most advanced Machine Learning algorithms out there

Data Science Goals



Use case: Mishandled Bags (MHB's)

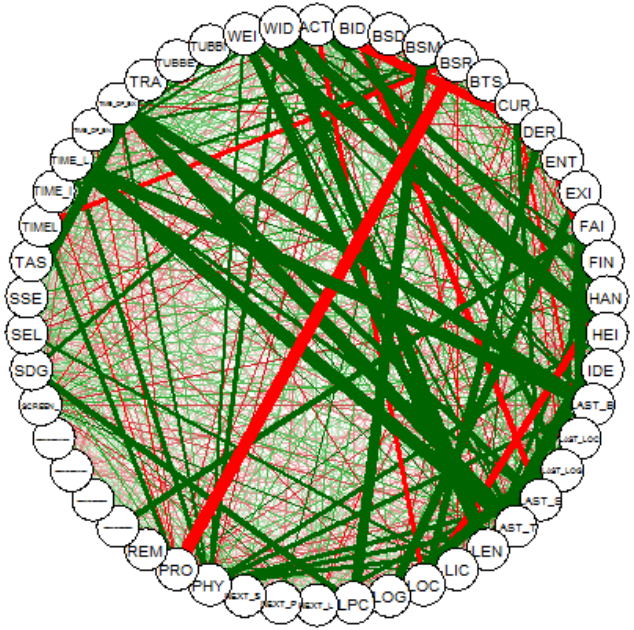
Use Case: Mishandled Bags (Mhb's)

- MHB's are bags that:
  - Are lost during handling
  - Miss their flight
  - Are put on the wrong airplane
- On a yearly basis the total cost of MHB's is **$2,4 bn**.
- Collaboration with a large airport and a manufacturer of baggage handling systems how to reduce these costs

The Data

Data of multiple sources in Splunk:

- Sample of over 30.000 unique bags
- 3 Month period
- 125 Variables for each bag, i.e:
  - Weight
  - Carrier
  - System entrypoint
  - Check-in to departure delta
  - Etc.
- 2% Was considered a MHB



The Approach

1. Use Splunk to combine and format the data
2. Use R to train a Boosted Decision Tree model, capable of performing Binary Logistic Classification (with 0=ok, 1=MHB)
3. Export the trained model as an R-package
4. Use the R-app to combine Splunk and R-code
5. Import the model and quickly classify new bags with a risk score
6. Extract the feature importance to assess which variables have the biggest impact on bags, causing MHB's

Risk Assessment

- The model returns a score between 0 and 1
- This is the probability of a bag becoming a MHB
- Use Splunk to filter low-risk bags
- Use Splunk alerting for high-risk bags

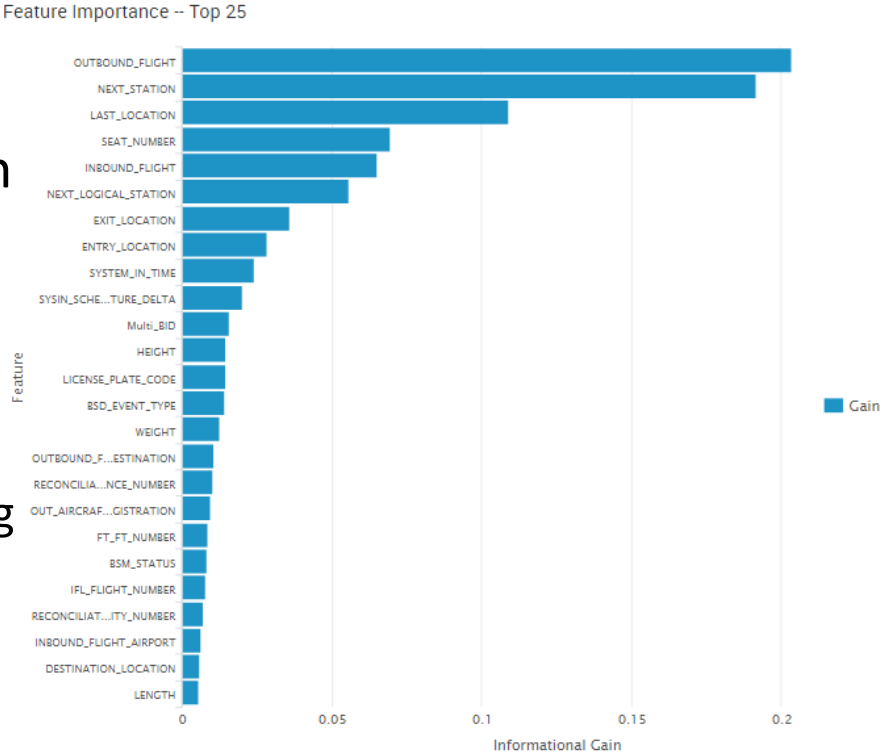
Bag Risk Assesment -- Risk > 5%

_time	LPC	Risk
2016-07-25 14:05:36	74363331	0.949320
2016-07-25 14:05:36	74919963	0.192906
2016-07-25 14:05:36	74698178	0.799526
2016-07-25 14:05:36	784716347	0.086178
2016-07-25 14:05:36	281689576	0.736813
2016-07-25 14:05:36	2141922645	0.973232
2016-07-25 14:05:36	2074538118	0.050471
2016-07-25 14:05:36	9352000179	0.953725
2016-07-25 14:05:36	74820613	0.110761
2016-07-25 14:05:36	784575182	0.066916
2016-07-25 14:05:36	74463944	0.093696
2016-07-25 14:05:36	7006297004	0.973772
2016-07-25 14:05:36	8006805224	0.327457
2016-07-25 14:05:36	7006649869	0.249184
2016-07-25 14:05:36	6006863284	0.915263
2016-07-25 14:05:36	74949557	0.821514
2016-07-25 14:05:36	74761598	0.086923
2016-07-25 14:05:36	2006226115	0.953627
2016-07-25 14:05:36	74310725	0.055793
2016-07-25 14:05:36	6006598858	0.087242

« prev 1 2 next »

Feature Importance

- "Reverse-engineered" the model to extract informational gain from each variable
- Several interesting finds:
  - Next station / Last location
  - Seat Number
  - Amount of Bag ID codes per unique bag
- Next steps: further investigate important variables with domain experts



Architecture And Challenges

- The R app on Spunkbase runs R on the Search Head
- Training models can be very CPU and Memory intensive
- Scaling a Search Head cluster becomes more complex

Architecture And Challenges

- The new R app communicates with a remote R server
- Training a model no longer impacts search performance
- Scale R independently from the Search Head cluster

The New App

- Is released TODAY!
- Contains several example dashboards and a new R code editor
- Contains new search commands for interacting with R
- Makes R an extention of SPL and allows you to create business value even faster

The New App

DEMO

R

Editor
Demo views

R app contents

This page is an overview of all the fun things you can do with R in Splunk.

Edit More Info

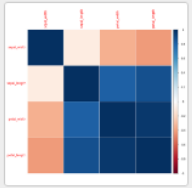
Filter

Editor

Script editor

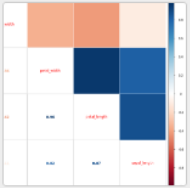
Run arbitrary R code on data from a Splunk search.

Demo views



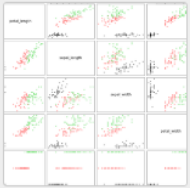
Correlation plot

Run the R corrplot command straight from a Splunk dashboard panel and display the graph.



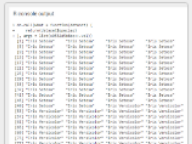
Correlation plot (mixed)

Run the R corrplot command straight from a dashboard panel and have different visualizations for top and bottom.



Pairs plot

Run the R pairs command straight from Splunk and display the graph in a dashboard panel.



Console output

Return the console output to a Splunk dashboard panel.

Demo view - Console output

Return the console output to a Splunk dashboard panel.

Edit More Info

R console output

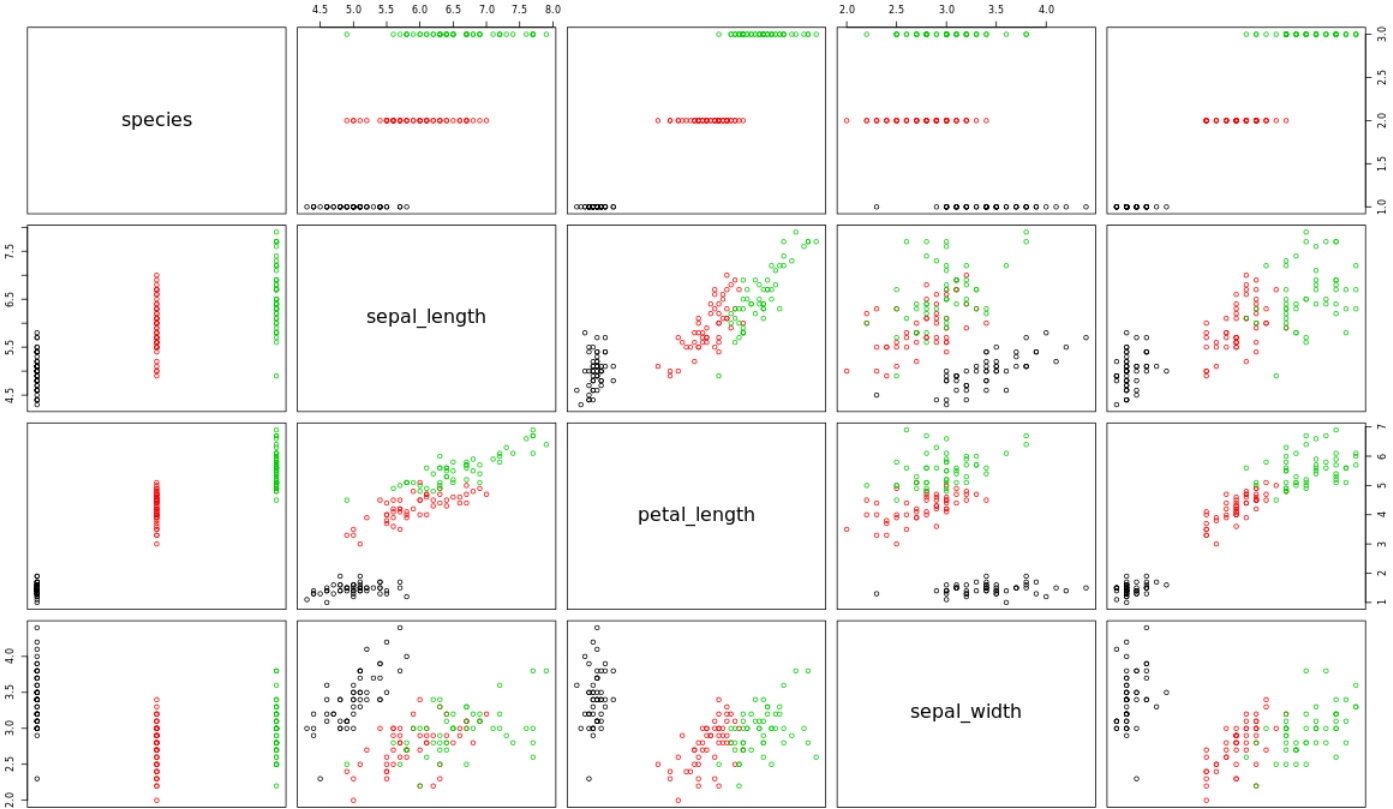
```
> do.call(args = list(x0362c35718::.val), what = function(dataset) {
+     return(dataset$species)
+ })
  [1] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
  [5] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
  [9] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [13] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [17] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [21] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [25] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [29] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [33] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [37] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [41] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [45] "Iris Setosa"     "Iris Setosa"     "Iris Setosa"     "Iris Setosa"
 [49] "Iris Setosa"     "Iris Setosa"     "Iris Versicolor" "Iris Versicolor"
 [53] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [57] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [61] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [65] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [69] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [73] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [77] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [81] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [85] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [89] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [93] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
 [97] "Iris Versicolor" "Iris Versicolor" "Iris Versicolor" "Iris Versicolor"
[101] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[105] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[109] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[113] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[117] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[121] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[125] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[129] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[133] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[137] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[141] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[145] "Iris Virginica"  "Iris Virginica"  "Iris Virginica"  "Iris Virginica"
[149] "Iris Virginica"  "Iris Virginica"
```

Demo view - Pairs

Run the R pairs command straight from Splunk and display the graph in a dashboard panel.

Edit More Info

R pairs command

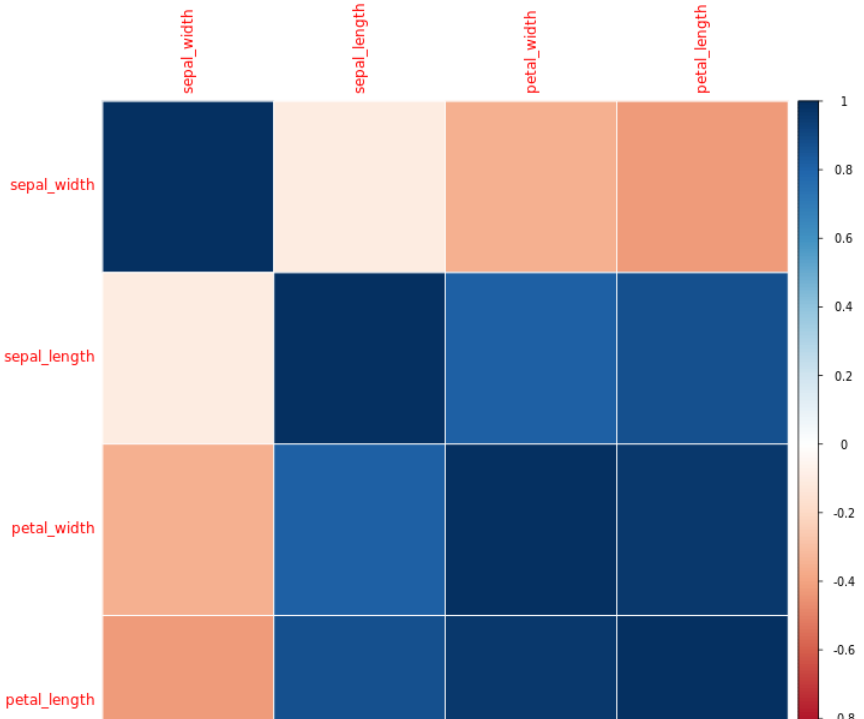


Demo view - Corrplot

Run the R corrplot command straight from a Splunk dashboard panel and display the graph.

Edit More Info

R correlation plot

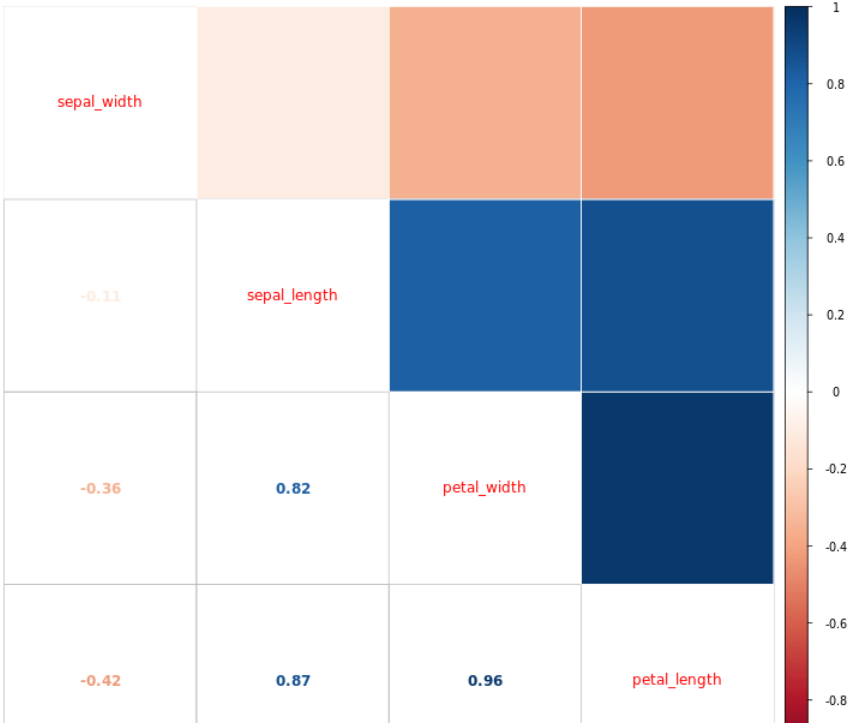


Demo view - Mixed corrplot

Run the R corrplot command straight from a dashboard panel and have different visualizations for top and bottom.

Edit More Info

R correlation plot mixed



Script editor

Run arbitrary R code on data from a Splunk search.

Edit | More Info

Query

```
| inputlookup iris.csv
```

All time

✓ 1 result (1/1/70 1:00:00.000 AM to 7/1/16 11:28:52.000 AM)

Job | Smart Mode

Script

```
# Convert the species column into a factor
dataset$species <- as.factor(dataset$species);

# Summarize the dataset
str(dataset);

# Visualize the relations
pairs(dataset, col = dataset$species);

# Print the entire dataset to console
print(dataset);
```

Run

Libraries

an

- AnomalyDetection
- Ckmeans.1d.dp
- lavaan
- pander
- quantreg
- randomForest
- translations

Functions

Filter

- AnomalyDetectionVec
- raw_data

AnomalyDetectionVec {AnomalyDetection} R Documentation

Anomaly Detection Using Seasonal Hybrid ESD Test

Description

A technique for detecting anomalies in seasonal univariate time series where the input is a series of observations.

Usage

```
AnomalyDetectionVec(x, max_anoms = 0.1, direction = "pos", alpha = 0.05,
  period = NULL, only_last = F, threshold = "None", e_value = F,
  longterm_period = NULL, plot = F, y_log = F, xlabel = "",
  ylabel = "count", title = NULL, verbose = FALSE)
```

Arguments

```
longterm_period = NULL, plot = F, y_log = F, xlabel = "",
ylabel = "count", title = NULL, verbose = FALSE)
```

Run

Arguments

Output

✓ 1 event (1/1/70 1:00:00.000 AM to 7/1/16 11:28:54.000 AM)

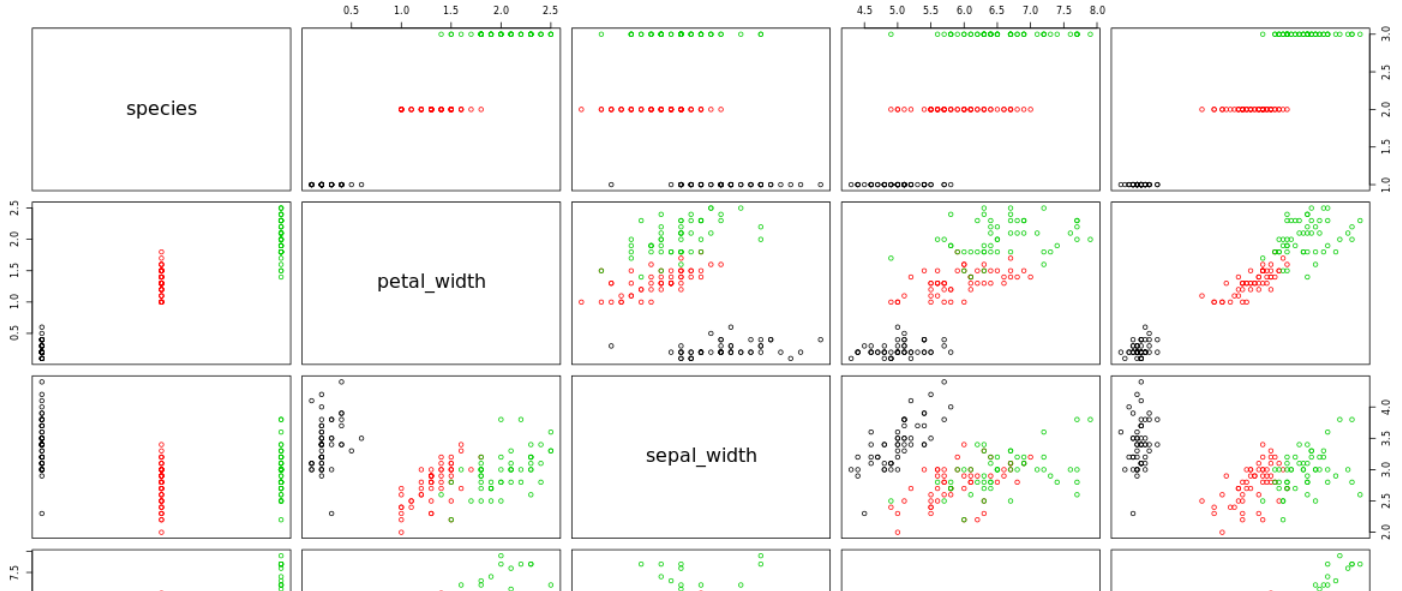
Job ⏸ ■ Smart Mode

Data Preview | R Console | R Graphics

R Graphics

Width 1200

Height 800



THANK YOU

splunk>