Case Study: Application of Meta-data and Best Practices to Scientific Data

Anderson; Swanson; Blank; Seiver; Allen; Gu; Yang and Griggs

October 4, 2014

1 Overview

This document exists as a Case Study: Application of Meta-data and Best Practices to Scientific Data. The team worked on it as part of Science Hack Day 2014 (San Francisco) - a 24 hour collaborative and hack event.

2 Objectives

The team had two primary objectives:
1) To augment information the core data, to help data consumers with interpretation
2) To encourage scientists to follow data 'best practices' when making data public

The team leveraged R, RAW, D3.js tools - and sought to demonstrate one method on how to impart better meta-data onto data
– Metadata Fields:
Data Points
column label
independent/dependent
continuous/categorical
unit of measurement (continuous data)
labels for categories (categorical data)
– Full Data Set
Author Name
Title
Date of Entry
Date(s) of Study
Location of Study
Overview
Purpose
Summary of data points

3 Data Origin and Data

Talk about data origin here - where and when it was gathered

4 Other Sources

reference other people or sources of information relevant to data

5 Data Summary

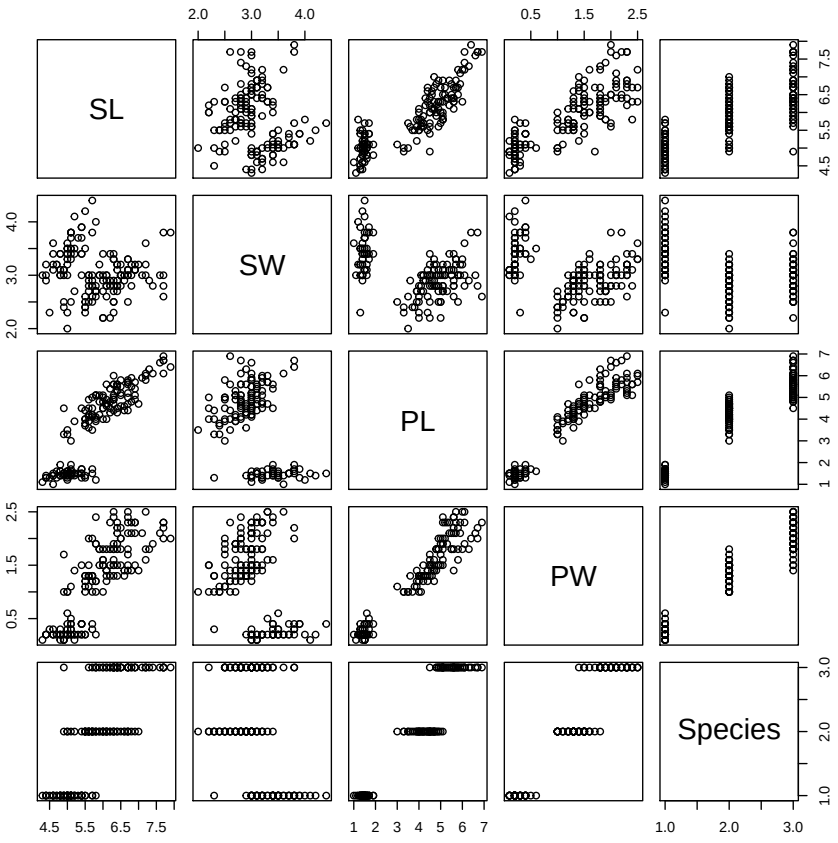
Below is the summary of count of ROWS and COLUMNS for the raw data, on the date the report was generated:

```
data <- read.csv("iris.CSV")
dim(data)  ### ROW COUNT and COLUMN COUNT

## [1] 150   5
```

```
summary(data)

##         SL              SW              PL              PW
##  Min.   :4.30   Min.   :2.00   Min.   :1.00   Min.   :0.1
```

```
##  1st Qu.:5.10   1st Qu.:2.80   1st Qu.:1.60   1st Qu.:0.3
##  Median :5.80   Median :3.00   Median :4.35   Median :1.3
##  Mean   :5.84   Mean   :3.05   Mean   :3.76   Mean   :1.2
##  3rd Qu.:6.40   3rd Qu.:3.30   3rd Qu.:5.10   3rd Qu.:1.8
##  Max.   :7.90   Max.   :4.40   Max.   :6.90   Max.   :2.5
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
plot(data)
```



```
REPORT <- read.csv("REPORT.csv")
print(REPORT[,2], max.levels = 0)

## [1] REPORT NAME: IRIS DATA SET
## [2] Version 0.01
## [3] REPORT DATE: 2014-10-04 23:50:23
## [4] AUTHOR: Ryan Anderson, Berkeley, California
## [5] test
## [6] this
## [7] that
```

6 Our Team

Science Hack Day 2014 Team Members:
Ryan Anderson
Chris Swanson
Jen Blank
Elizabeth Seiver
David Allen
Angela Gu
Anderson Yang
Kelly Griggs
and thanks to all the open source tools like R, knitr, D3 (Bostock), and RAW authors.