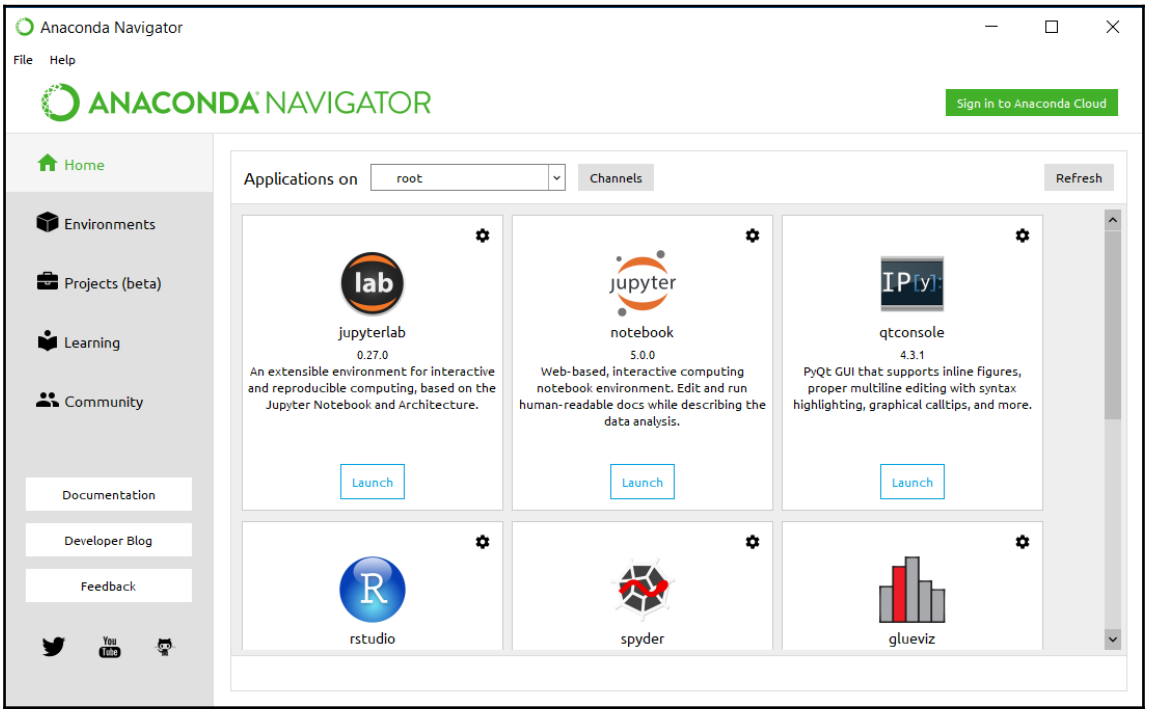
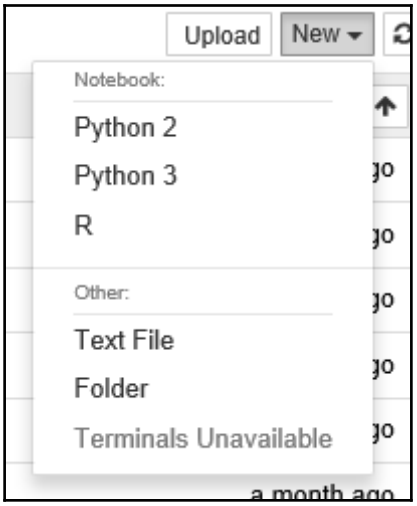
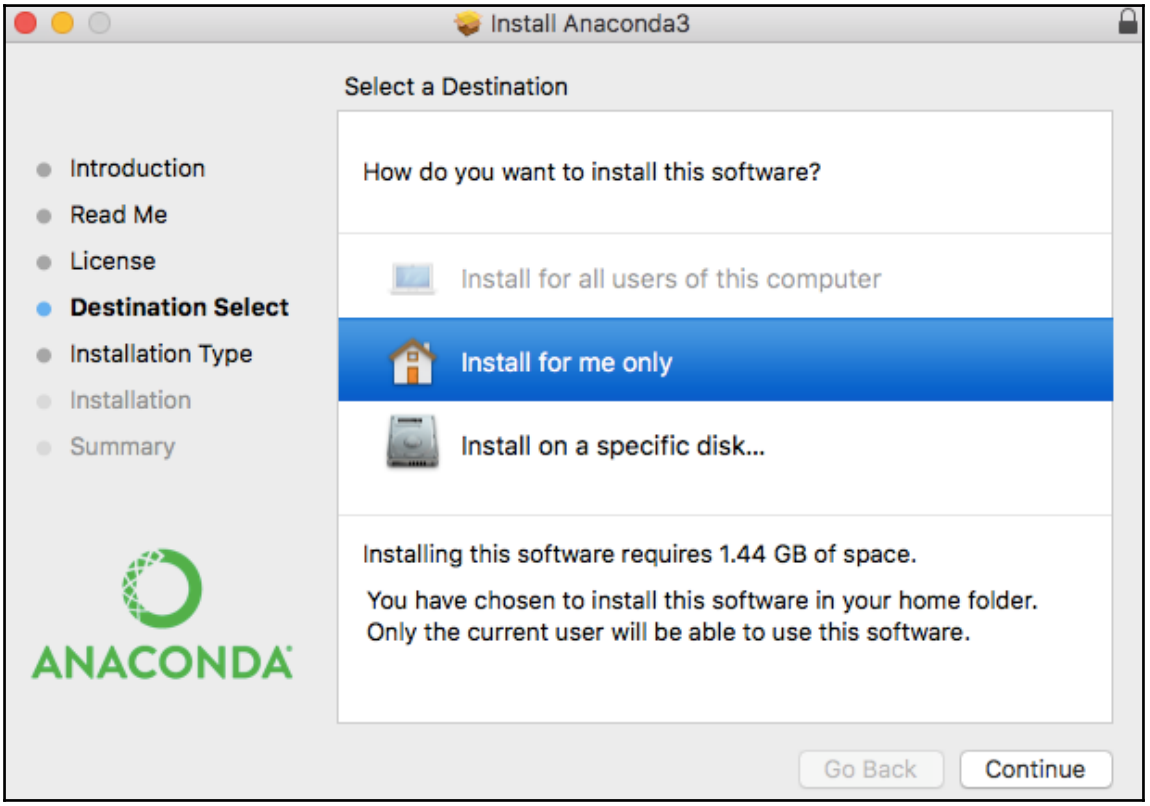
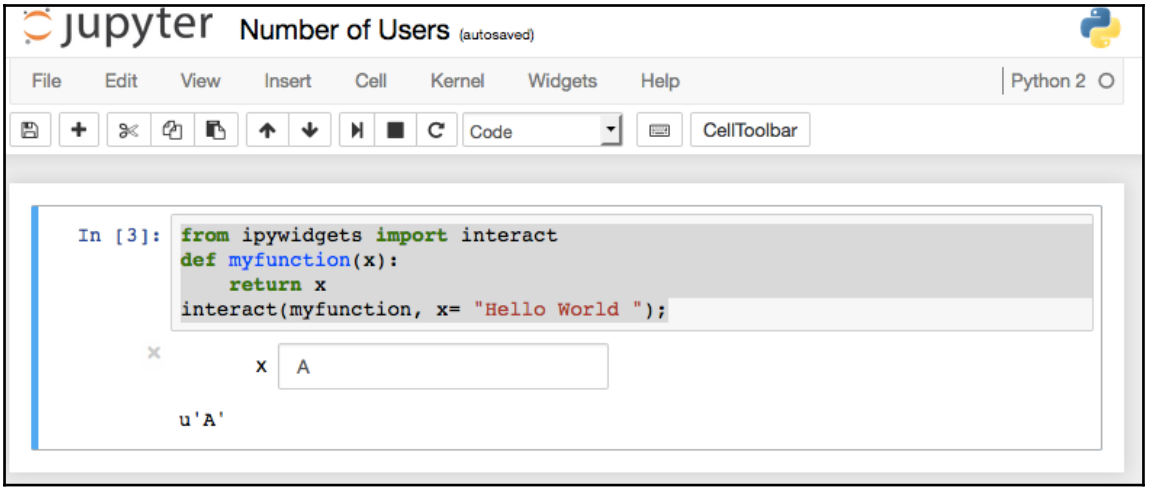
Chapter 1: Installation and Setting up the Environment

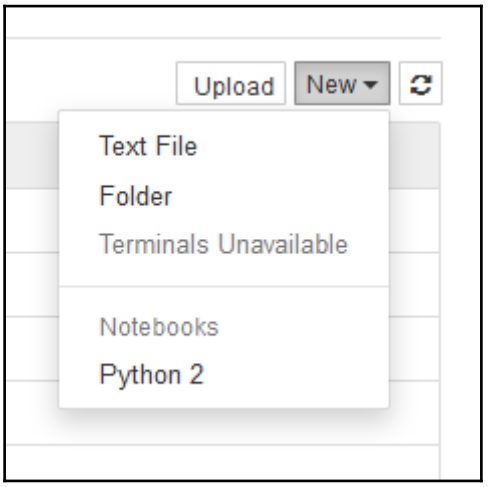


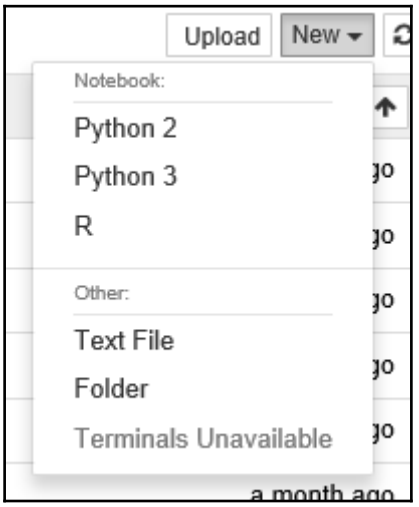

Upload
New
Notebook:
Python 2
Python 3
R
Other:
Text File
Folder
Terminals Unavailable

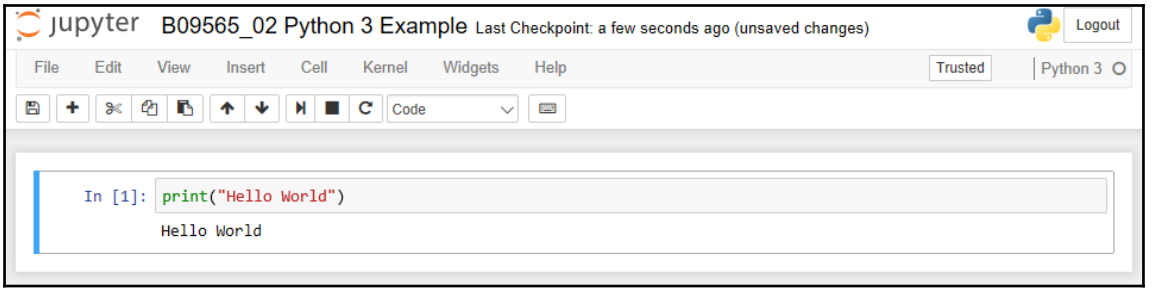

Install Anaconda3
Select a Destination
Introduction
Read Me
License
Destination Select
Installation Type
Installation
Summary
How do you want to install this software?
Install for all users of this computer
Install for me only
Install on a specific disk...
Installing this software requires 1.44 GB of space.
You have chosen to install this software in your home folder.
Only the current user will be able to use this software.
ANACONDA
Go Back
Continue

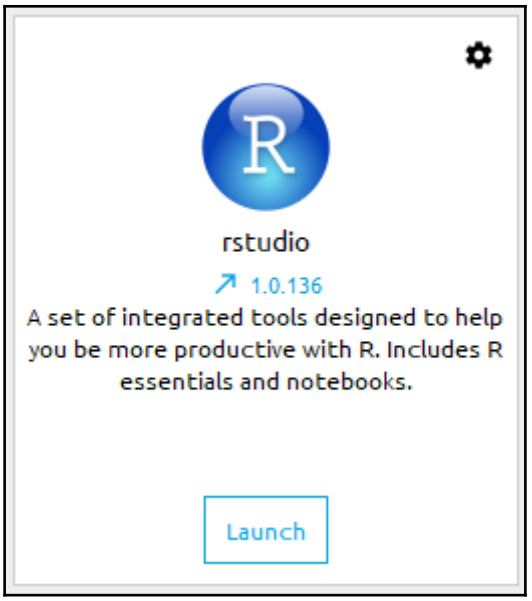

jupyter
Number of Users (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python 2
Code
CellToolbar
In [3]:
from ipywidgets import interact
def myfunction(x):
return x
interact(myfunction, x= "Hello World ");
x
A
u'A'

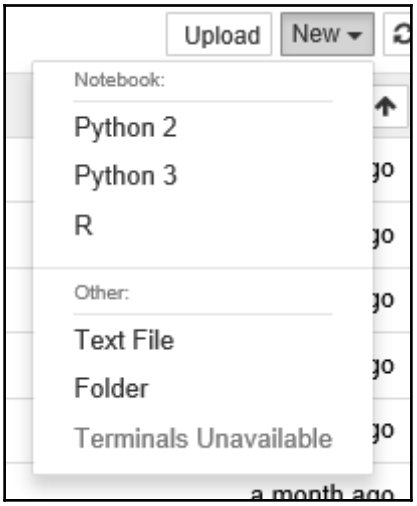
Chapter 2: Adding an Engine

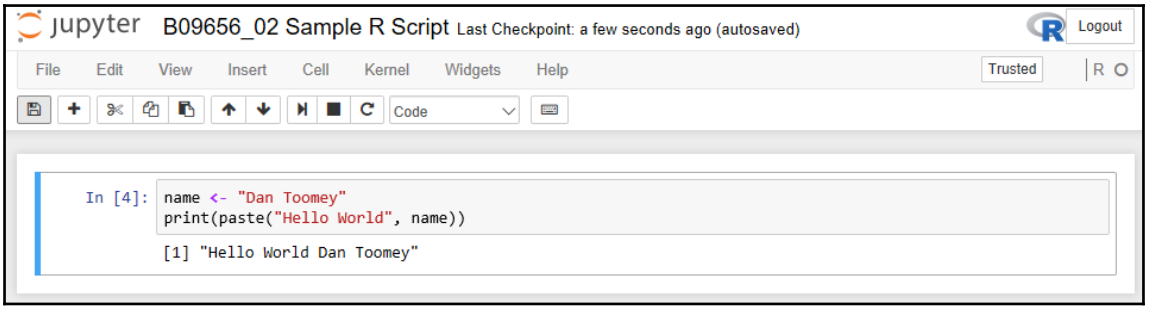


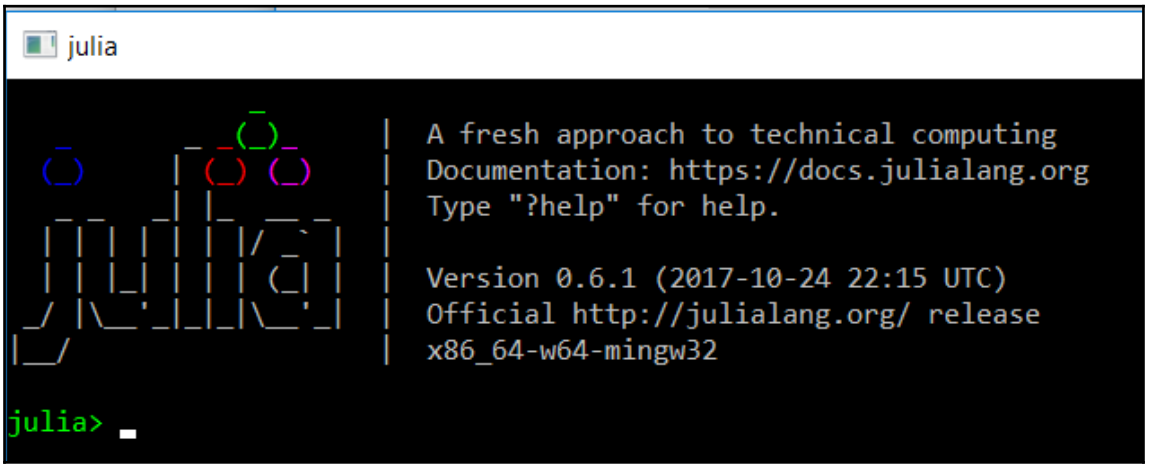



jupyter
B09565_02 Python 3 Example
Last Checkpoint: a few seconds ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Code
In [1]:
print("Hello World")
Hello World

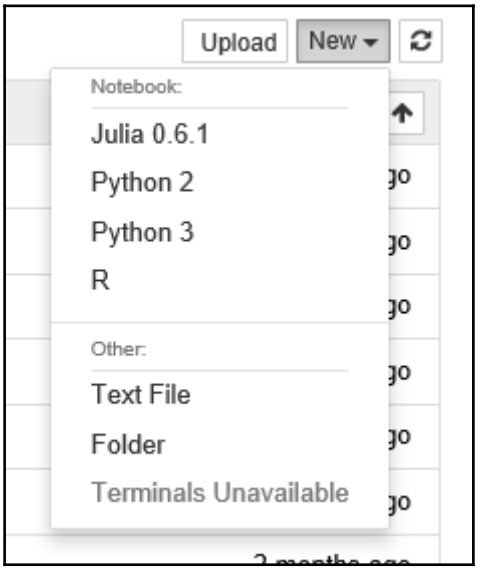

rstudio
1.0.136
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.
Launch


Upload
New
Notebook:
Python 2
Python 3
R
Other:
Text File
Folder
Terminals Unavailable
a month ago


jupyter
B09656_02 Sample R Script
Last Checkpoint: a few seconds ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
R
Code
In [4]:
name <- "Dan Toomey"
print(paste("Hello World", name))
[1] "Hello World Dan Toomey"


julia
A fresh approach to technical computing
Documentation: https://docs.julialang.org
Type "?help" for help.
Version 0.6.1 (2017-10-24 22:15 UTC)
Official http://julialang.org/ release
x86_64-w64-mingw32
julia>

```
julia> Pkg.add("IJulia")
INFO: Initializing package repository C:\Users\Dan\.julia\v0.6
INFO: Cloning METADATA from https://github.com/JuliaLang/METADATA.jl
INFO: Cloning cache of BinDeps from https://github.com/JuliaLang/BinDeps.jl.git
INFO: Cloning cache of BufferedStreams from https://github.com/BioJulia/BufferedStreams.jl.git
INFO: Cloning cache of Compat from https://github.com/JuliaLang/Compat.jl.git
INFO: Cloning cache of Conda from https://github.com/JuliaPy/Conda.jl.git
INFO: Cloning cache of IJulia from https://github.com/JuliaLang/IJulia.jl.git
INFO: Cloning cache of JSON from https://github.com/JuliaIO/JSON.jl.git
INFO: Cloning cache of LibExpat from https://github.com/JuliaIO/LibExpat.jl.git
INFO: Cloning cache of Libz from https://github.com/BioJulia/Libz.jl.git
INFO: Cloning cache of MbedTLS from https://github.com/JuliaWeb/MbedTLS.jl.git
INFO: Cloning cache of SHA from https://github.com/staticfloat/SHA.jl.git
INFO: Cloning cache of URIParser from https://github.com/JuliaWeb/URIParser.jl.git
INFO: Cloning cache of WinRPM from https://github.com/JuliaPackaging/WinRPM.jl.git
INFO: Cloning cache of ZMQ from https://github.com/JuliaInterop/ZMQ.jl.git
INFO: Installing BinDeps v0.7.0
```



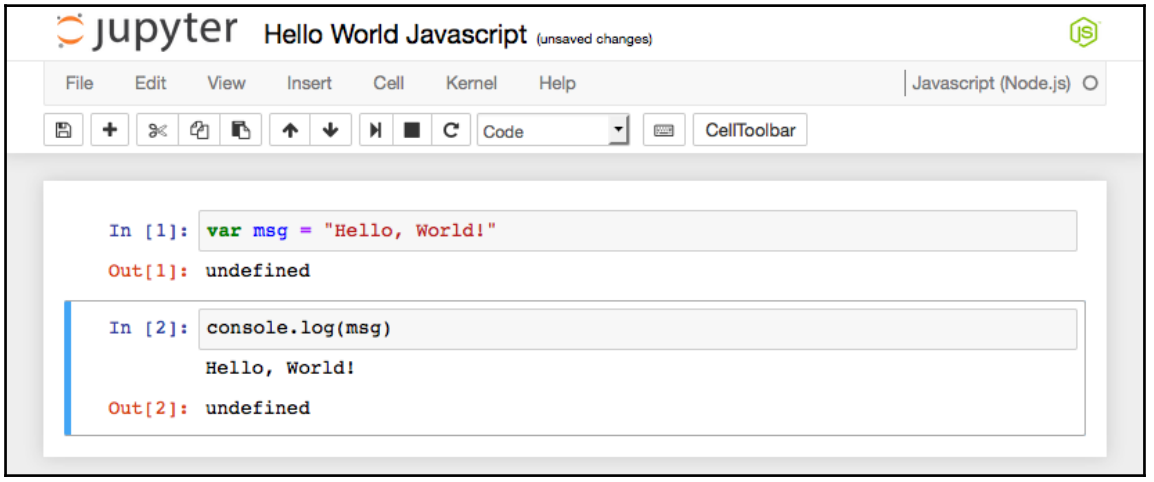
In [2]:

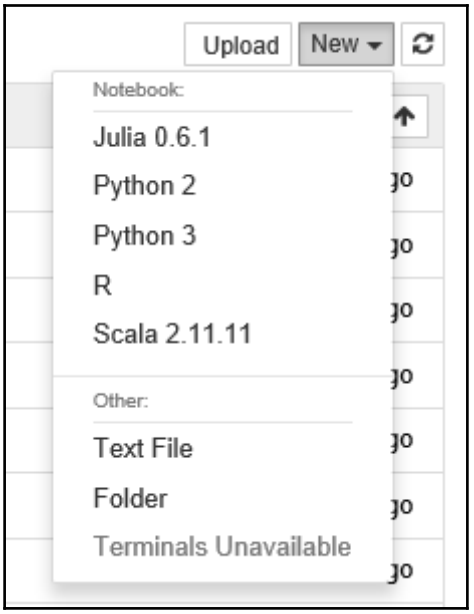
```
Pkg.add("RDatasets")
using RDatasets
describe(dataset("datasets", "iris"))
```

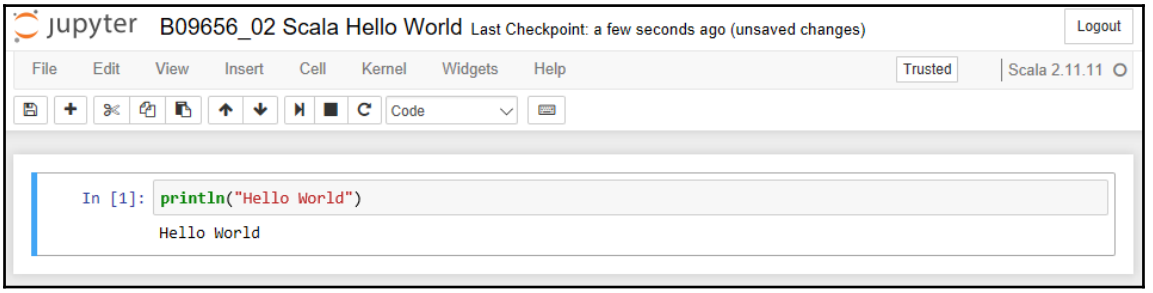
```
INFO: Cloning cache of DataArrays from https://github.com/Julia
INFO: Cloning cache of DataFrames from https://github.com/Julia
INFO: Cloning cache of DataStructures from https://github.com/J
INFO: Cloning cache of FileIO from https://github.com/JuliaIO/F
INFO: Cloning cache of GZip from https://github.com/JuliaIO/GZi
INFO: Cloning cache of RData from https://github.com/JuliaStats
INFO: Cloning cache of RDatasets from https://github.com/johnmy
INFO: Cloning cache of Reexport from https://github.com/simonst
INFO: Cloning cache of SortingAlgorithms from https://github.co
INFO: Cloning cache of SpecialFunctions from https://github.com
INFO: Cloning cache of StatsBase from https://github.com/JuliaS
INFO: Installing DataArrays v0.6.2
INFO: Installing DataFrames v0.10.1
INFO: Installing DataStructures v0.7.2
INFO: Installing FileIO v0.5.2
INFO: Installing GZip v0.3.0
INFO: Installing RData v0.1.0
INFO: Installing RDatasets v0.2.0
INFO: Installing Reexport v0.0.3
INFO: Installing SortingAlgorithms v0.1.1
INFO: Installing SpecialFunctions v0.3.4
INFO: Installing StatsBase v0.19.0
INFO: Building SpecialFunctions
INFO: Package database updated
INFO: Precompiling module Reexport.
INFO: Precompiling module FileIO.
INFO: Precompiling module DataFrames.
INFO: Precompiling module RData.
```

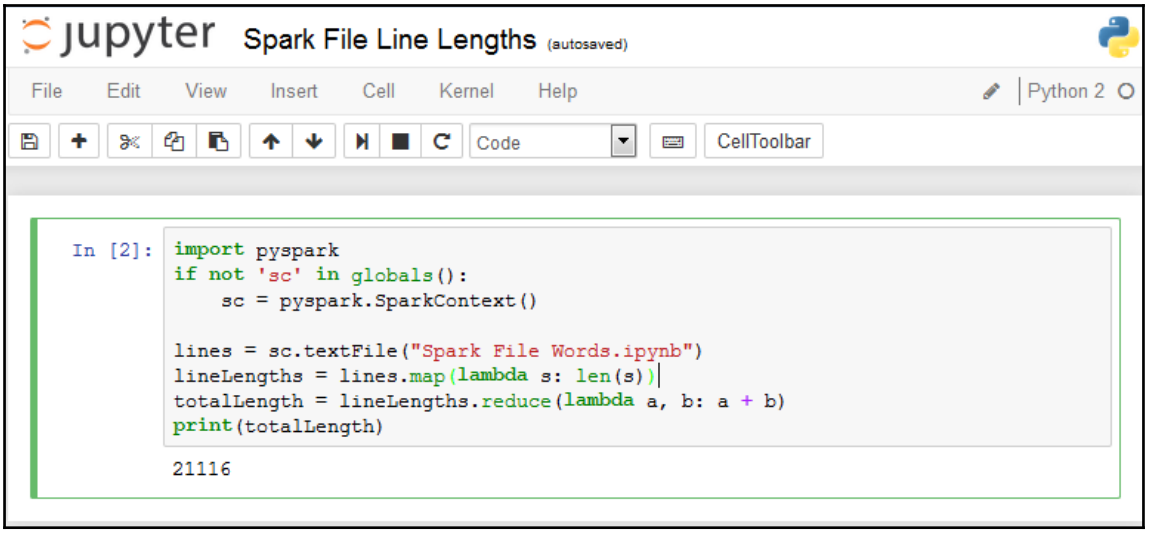
```
SepalLength
Summary Stats:
Mean:           5.843333
Minimum:        4.300000
1st Quartile:   5.100000
Median:         5.800000
3rd Quartile:   6.400000
Maximum:        7.900000
Length:         150
Type:           Float64
Number Missing: 0
% Missing:      0.000000
```

Upload
New
Notebook:
Javascript (Node.js)
Julia 0.6.1
Python 2
Python 3
R
Scala 2.11.11
Other:
Text File
Folder
Terminals Unavailable
a month ago


jupyter
Hello World Javascript
(unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Help
Javascript (Node.js)
Code
CellToolbar
In [1]: var msg = "Hello, World!"
Out[1]: undefined
In [2]: console.log(msg)
Hello, World!
Out[2]: undefined


Upload
New
Notebook:
Julia 0.6.1
Python 2
Python 3
R
Scala 2.11.11
Other:
Text File
Folder
Terminals Unavailable


jupyter
B09656_02 Scala Hello World
Last Checkpoint: a few seconds ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Scala 2.11.11
Code
In [1]:
println("Hello World")
Hello World


jupyter
Spark File Line Lengths (autosaved)
File Edit View Insert Cell Kernel Help
Python 2
Code
CellToolbar
In [2]:
import pyspark
if not 'sc' in globals():
sc = pyspark.SparkContext()
lines = sc.textFile("Spark File Words.ipynb")
lineLengths = lines.map(lambda s: len(s))
totalLength = lineLengths.reduce(lambda a, b: a + b)
print(totalLength)
21116

Chapter 3: Accessing and Retrieving Data

```
In [2]: head(heating)
```

X	idcase	depvar	ic.gc	ic.gr	ic.ec	ic.er	ic.hp	oc.gc	oc.gr	...	oc.hp	income	agehed
1	1	gc	866.00	962.64	859.90	995.76	1135.50	199.69	151.72	...	237.88	7	25
2	2	gc	727.93	758.89	796.82	894.69	968.90	168.66	168.66	...	199.19	5	60
3	3	gc	599.48	783.05	719.86	900.11	1048.30	165.58	137.80	...	171.47	4	65
4	4	er	835.17	793.06	761.25	831.04	1048.70	180.88	147.14	...	222.95	2	50
5	5	er	755.59	846.29	858.86	985.64	883.05	174.91	138.90	...	178.49	2	25
6	6	gc	666.11	841.71	693.74	862.56	859.18	135.67	140.97	...	209.27	6	65

```
In [7]: head(heating)
```

X	system	install_cost	annual_cost	income	agehed	rooms	region	ratio_annual_install
1	gc	866.00	199.69	7	25	6	ncostl	4.336722
2	gc	727.93	168.66	5	60	5	scostl	4.315961
3	gc	599.48	165.58	4	65	2	ncostl	3.620486
4	er	835.17	180.88	2	50	4	scostl	4.617260
5	er	755.59	174.91	2	25	6	valley	4.319879
6	gc	666.11	135.67	6	65	7	scostl	4.909781

```
{
    "Models": [
        {
            "model_name": "021 C",
            "model_make_id": "ford"
        },
        {
            "model_name": "12 M",
            "model_make_id": "ford"
        },
```


```
//load the JSON dataset
//http://www.carqueryapi.com/api/0.3/?callback=?&cmd=getModels&make=ford
var fords = require('/Users/dtoomey/fords.json');

//display how many Ford models are in our data set
console.log("There are " + fords.Models.length + " Ford models in the data set");

//loop over the set
var index = 1
for(var i=0; i<fords.Models.length; i++) {

    //get this model
    var model = fords.Models[i];

    //pull it's name
    var name = model.model_name;

    //if the model name does not have numerics in it
    if(! name.match(/[0-9]/i)) {
        //display the model name
        console.log("Model " + index + " is a " + name);
        index++;
    }

    //only display the first 5
    if (index>5) break;
}
```

```
            There are 147 Ford models in the data set
            Model 1 is a Aerostar
            Model 2 is a Anglia
            Model 3 is a Artic
            Model 4 is a Aspire
            Model 5 is a Bantam

Out[27]: 5
```

```
 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
      Species
setosa    :50
versicolor:50
virginica :50
```

```
# Source:   table<iris> [?? x 5]
# Database: sqlite 3.19.3 []
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
          <dbl>       <dbl>        <dbl>       <dbl>   <chr>
 1          5.1         3.5          1.4         0.2  setosa
 2          4.9         3.0          1.4         0.2  setosa
 3          4.7         3.2          1.3         0.2  setosa
 4          4.6         3.1          1.5         0.2  setosa
 5          5.0         3.6          1.4         0.2  setosa
 6          5.4         3.9          1.7         0.4  setosa
 7          4.6         3.4          1.4         0.3  setosa
 8          5.0         3.4          1.5         0.2  setosa
 9          4.4         2.9          1.4         0.2  setosa
10          4.9         3.1          1.5         0.1  setosa
# ... with more rows
```

```
In [21]: head(iris_db, n = 10)
```

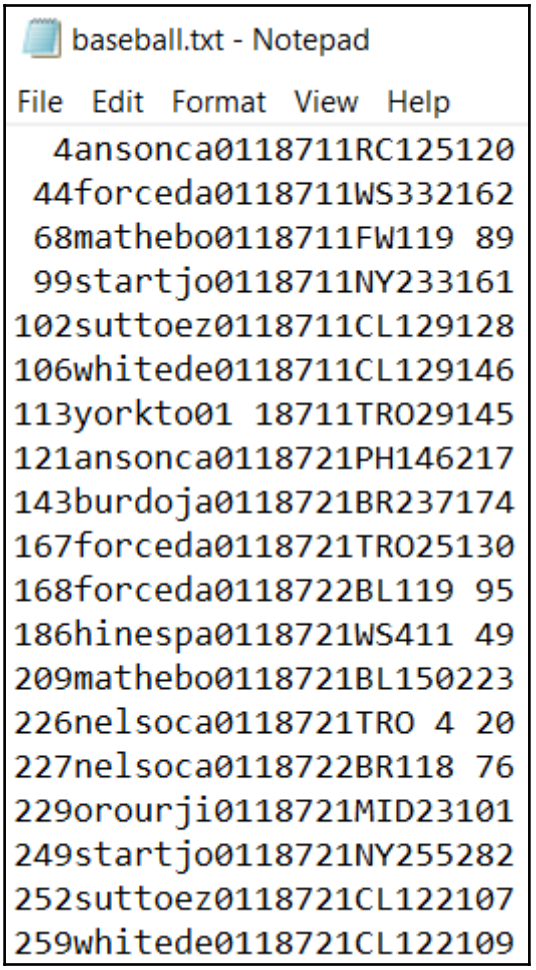
```
# Source:   lazy query [?? x 5]
# Database: sqlite 3.19.3 []
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
          <dbl>       <dbl>        <dbl>       <dbl>   <chr>
 1          5.1         3.5          1.4         0.2  setosa
 2          4.9         3.0          1.4         0.2  setosa
 3          4.7         3.2          1.3         0.2  setosa
 4          4.6         3.1          1.5         0.2  setosa
 5          5.0         3.6          1.4         0.2  setosa
 6          5.4         3.9          1.7         0.4  setosa
 7          4.6         3.4          1.4         0.3  setosa
 8          5.0         3.4          1.5         0.2  setosa
 9          4.4         2.9          1.4         0.2  setosa
10          4.9         3.1          1.5         0.1  setosa
# ... with more rows
```

```
In [19]: show_query(head(iris_db, n = 10))
```

```
<SQL>
SELECT *
FROM `iris`
LIMIT 10
```

```
In [20]: # pull data back into R using collect
my_iris <- iris_db %>% collect()
my_iris
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa


baseball.txt - Notepad
File Edit Format View Help
4ansonca0118711RC125120
44forceda0118711WS332162
68mathebo0118711FW119 89
99startjo0118711NY233161
102suttoez0118711CL129128
106whitede0118711CL129146
113yorkto01 18711TRO29145
121ansonca0118721PH146217
143burdoja0118721BR237174
167forceda0118721TRO25130
168forceda0118722BL119 95
186hinespa0118721WS411 49
209mathebo0118721BL150223
226nelsoca0118721TRO 4 20
227nelsoca0118722BR118 76
229orourji0118721MID23101
249startjo0118721NY255282
252suttoez0118721CL122107
259whitede0118721CL122109

Out[1]:

	row	id	year	stint	team	g	ab
0	4	ansonca01	1871	1	RC1	25	120
1	44	forceda01	1871	1	WS3	32	162
2	68	mathebo01	1871	1	FW1	19	89
3	99	startjo01	1871	1	NY2	33	161
4	102	suttoez01	1871	1	CL1	29	128
5	106	whitede01	1871	1	CL1	29	146
6	113	yorkto01	1871	1	TRO	29	145
7	121	ansonca01	1872	1	PH1	46	217
8	143	burdoja01	1872	1	BR2	37	174
9	167	forceda01	1872	1	TRO	25	130
10	168	forceda01	1872	2	BL1	19	95
11	186	hinespa01	1872	1	WS4	11	49
12	209	mathebo01	1872	1	BL1	50	223
13	226	nelsoca01	1872	1	TRO	4	20
14	227	nelsoca01	1872	2	BR1	18	76
15	229	orourji01	1872	1	MID	23	101
16	249	startjo01	1872	1	NY2	55	282
17	252	suttoez01	1872	1	CL1	22	107
18	259	whitede01	1872	1	CL1	22	109

```
  Length Class             Mode
1 2      PlainTextDocument list
2 2      PlainTextDocument list
3 2      PlainTextDocument list
4 2      PlainTextDocument list
5 2      PlainTextDocument list
```

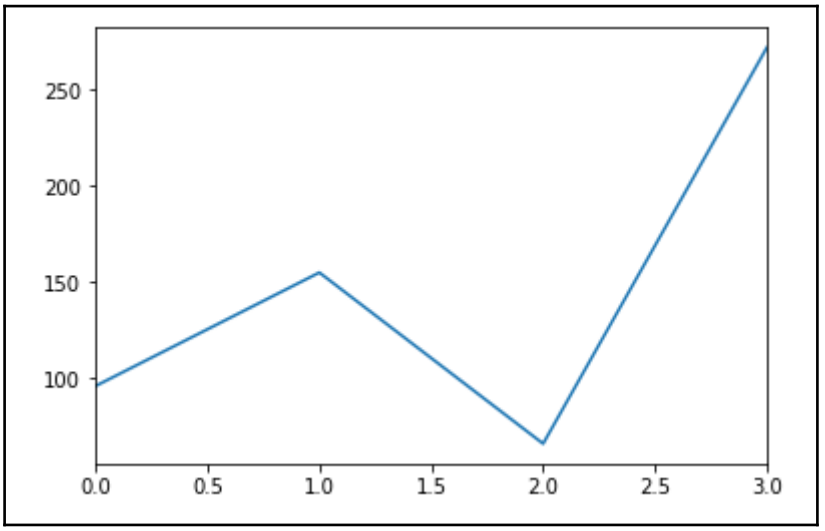
```
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] For the second time in a month,<U+00A0>President Trump discussed his plans to change the tax cod
e. This is an annotated transcript of his remarks, which took place at the Indiana Farm Bureau Buildi
ng.<U+00A0>Click on the highlighted statements below to see a brief summary of statements that we and
others have fact-checked.<U+00A0>For more information, please read our story, <U+0093><U+0091>Death T
ax<U+0092> Talking Point Won<U+0092>t Die.<U+0094>
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] for the second time in a month, president trump discussed his plans to change the tax code. this
is an annotated transcript of his remarks, which took place at the indiana farm bureau building. clic
k on the highlighted statements below to see a brief summary of statements that we and others have fa
ct-checked. for more information, please read our story, <U+0093><U+0091>death tax<U+0092> talking po
int won<U+0092>t die.<U+0094>
```

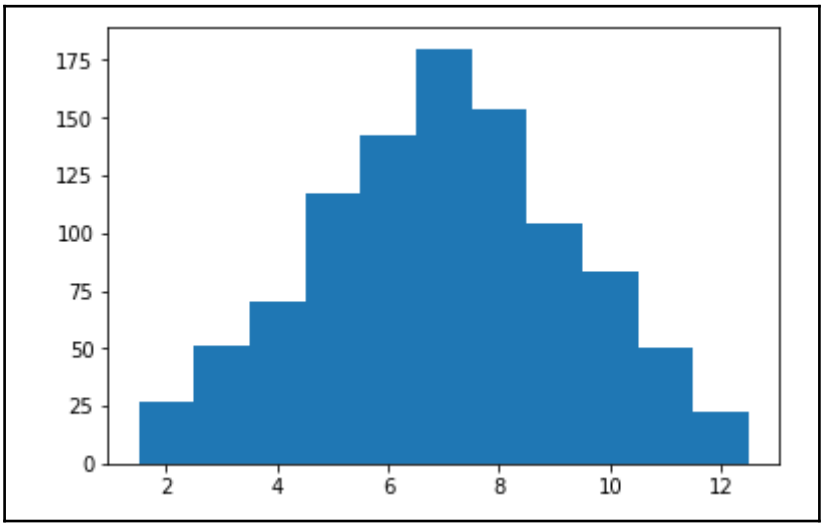
```
<<DocumentTermMatrix (documents: 94, terms: 1048)>>
Non-/sparse entries: 3092/95420
Sparsity           : 97%
Maximal term length: 15
Weighting          : term frequency (tf)
Sample             :
    Terms
Docs and applause for have our tax that the will you
  10   2        1   1    2   2   1    1   7    0   0
  11  11        0   0    3   0   0    3   6    0   1
  45   7        1   3    3   0   3    4   8    0   1
  47   2        2   1    0   0   2    4   2    0   0
  51   4        1   2    1   2   0    1   1    0   1
  64   4        2   1    0   1   0    1   5    2   0
  67   2        1   1    2   0   0    2   3    1   3
  73   5        1   3    0   2   1    0   4    0   0
  80   5        1   0    3   0   3    0   6    2   2
  9    5        0   0    3   2   0    4   4    0   1
```

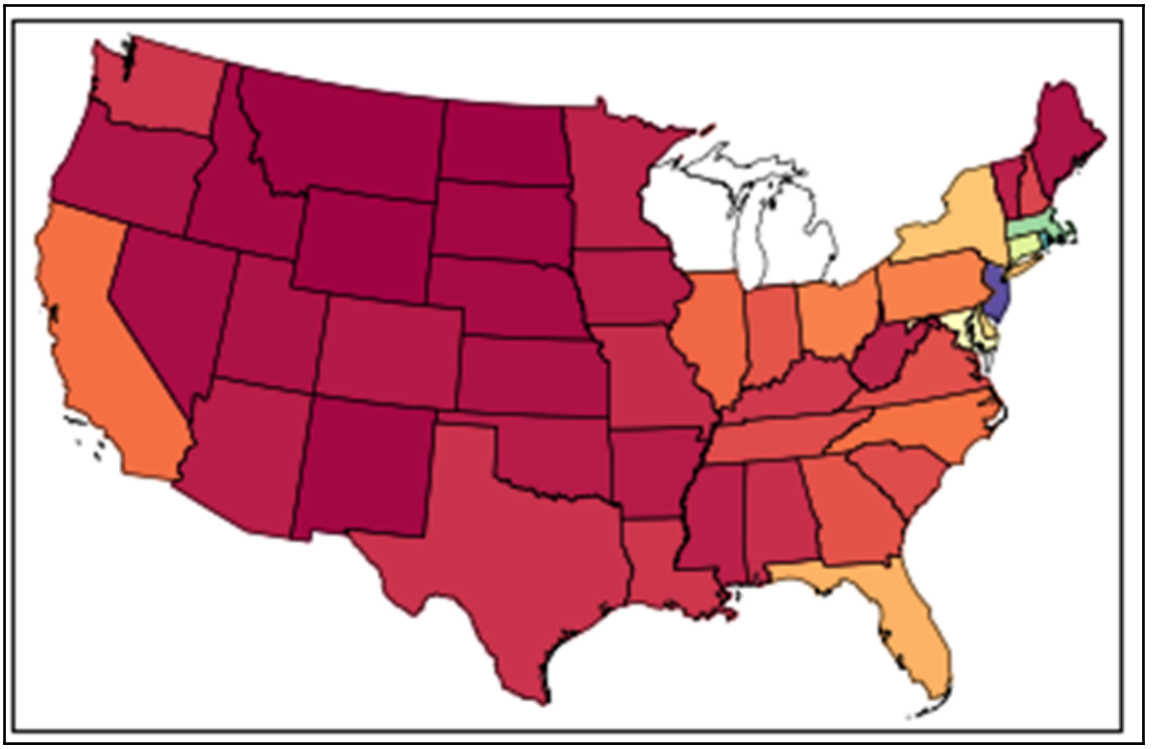
$tax =

the	0.44
cuts	0.43
pay	0.37
system	0.32
lower	0.3
significant	0.3
breaks	0.3
farms	0.3
competitors	0.3
numerous	0.3
death	0.29
tremendous	0.29
small	0.29
because	0.27
reform	0.27
calculate	0.27

Chapter 4: Visualizing Your Analytics

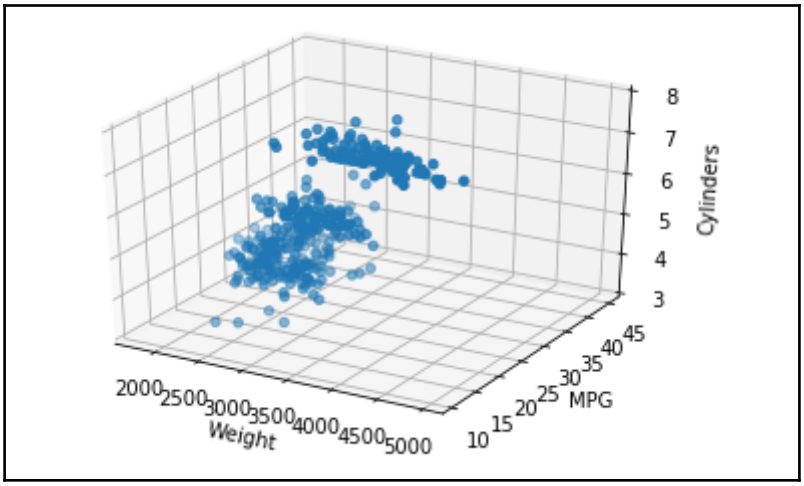


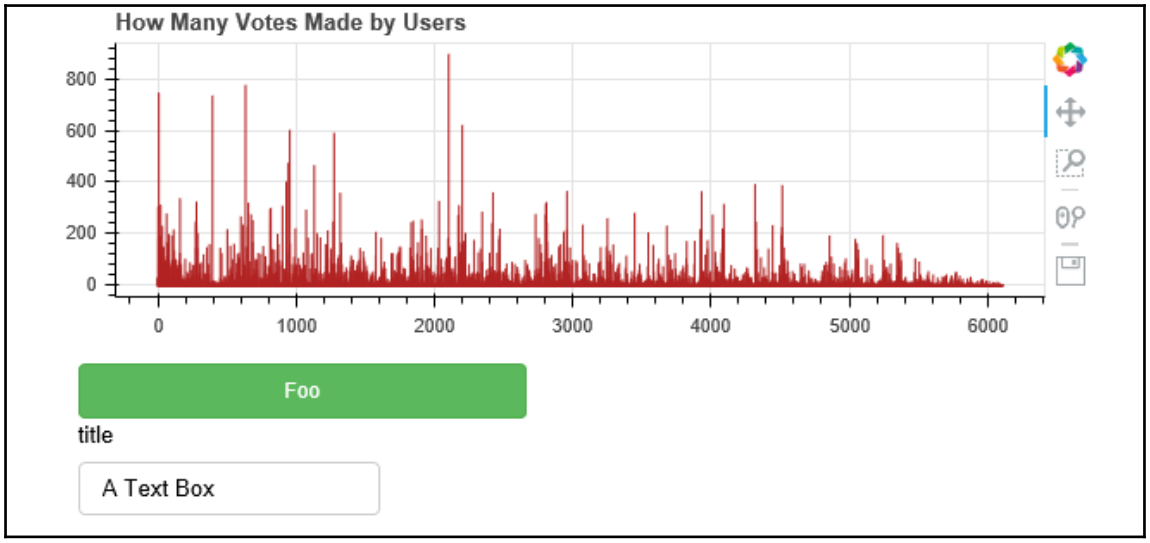




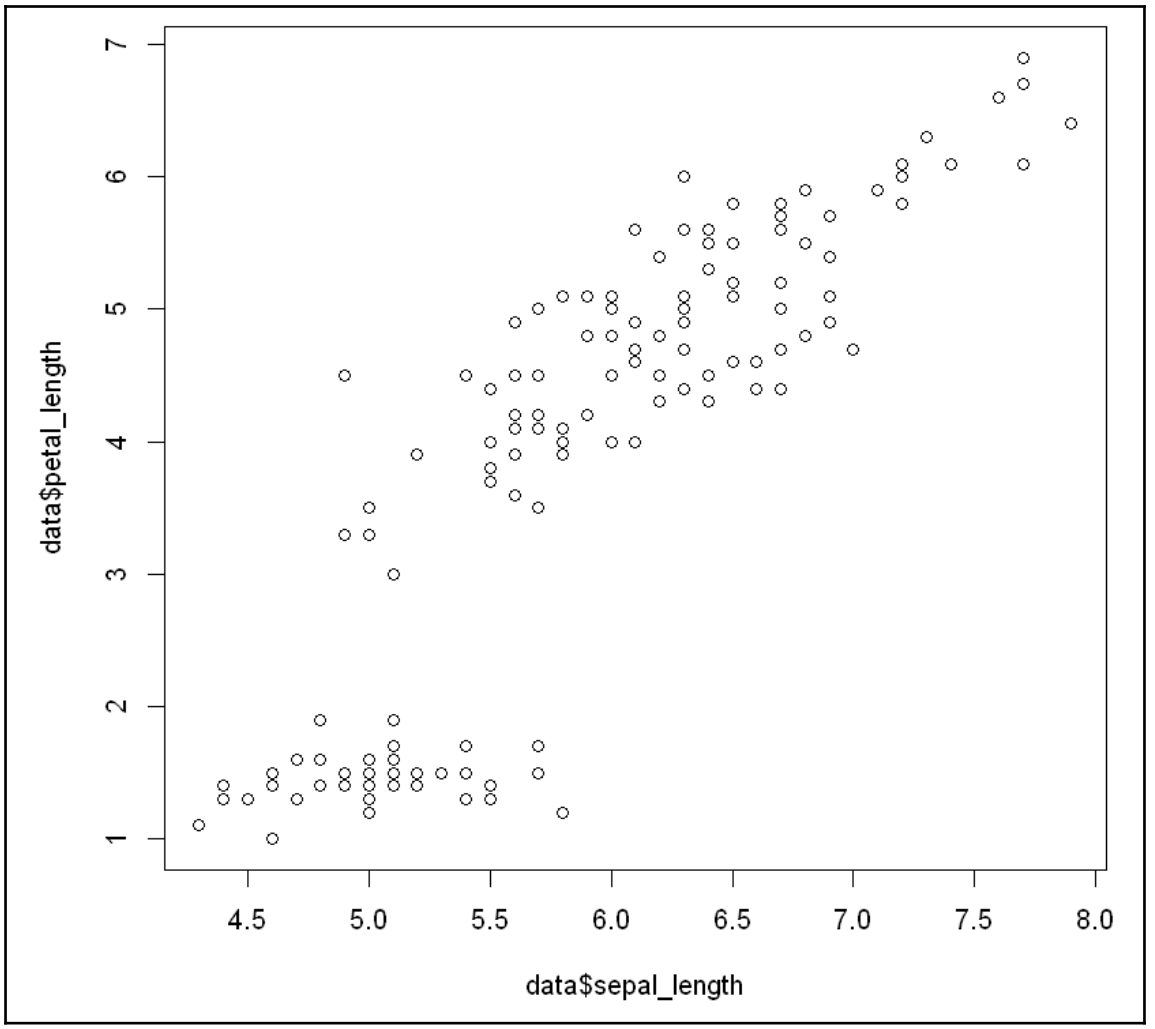
mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	\
18.0	8	307.0	130.0	3504.0	12.0	70	1	
15.0	8	350.0	165.0	3693.0	11.5	70	1	
18.0	8	318.0	150.0	3436.0	11.0	70	1	
16.0	8	304.0	150.0	3433.0	12.0	70	1	
17.0	8	302.0	140.0	3449.0	10.5	70	1	

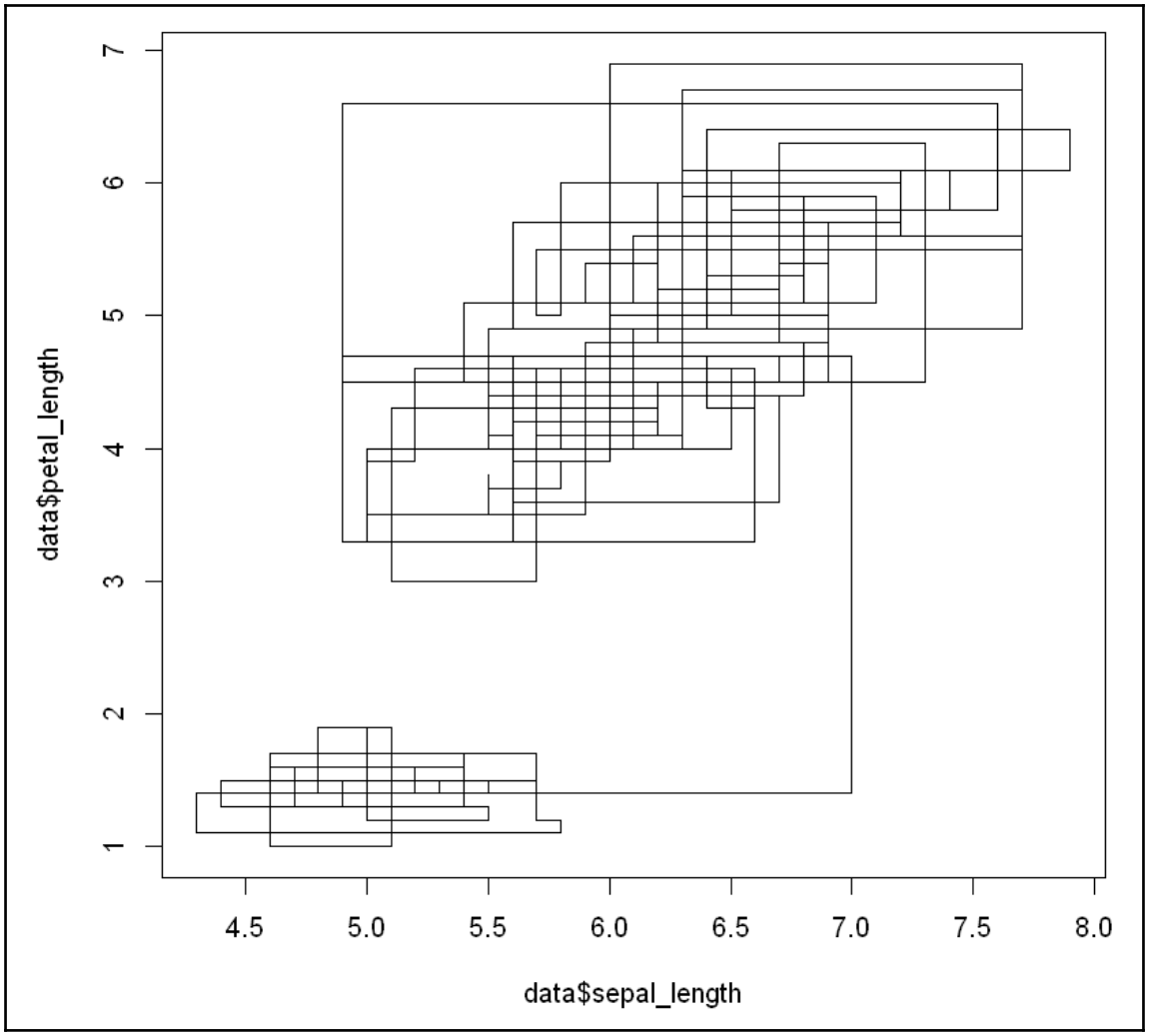
mpg	name
18.0	chevrolet chevelle malibu
15.0	buick skylark 320
18.0	plymouth satellite
16.0	amc rebel sst
17.0	ford torino

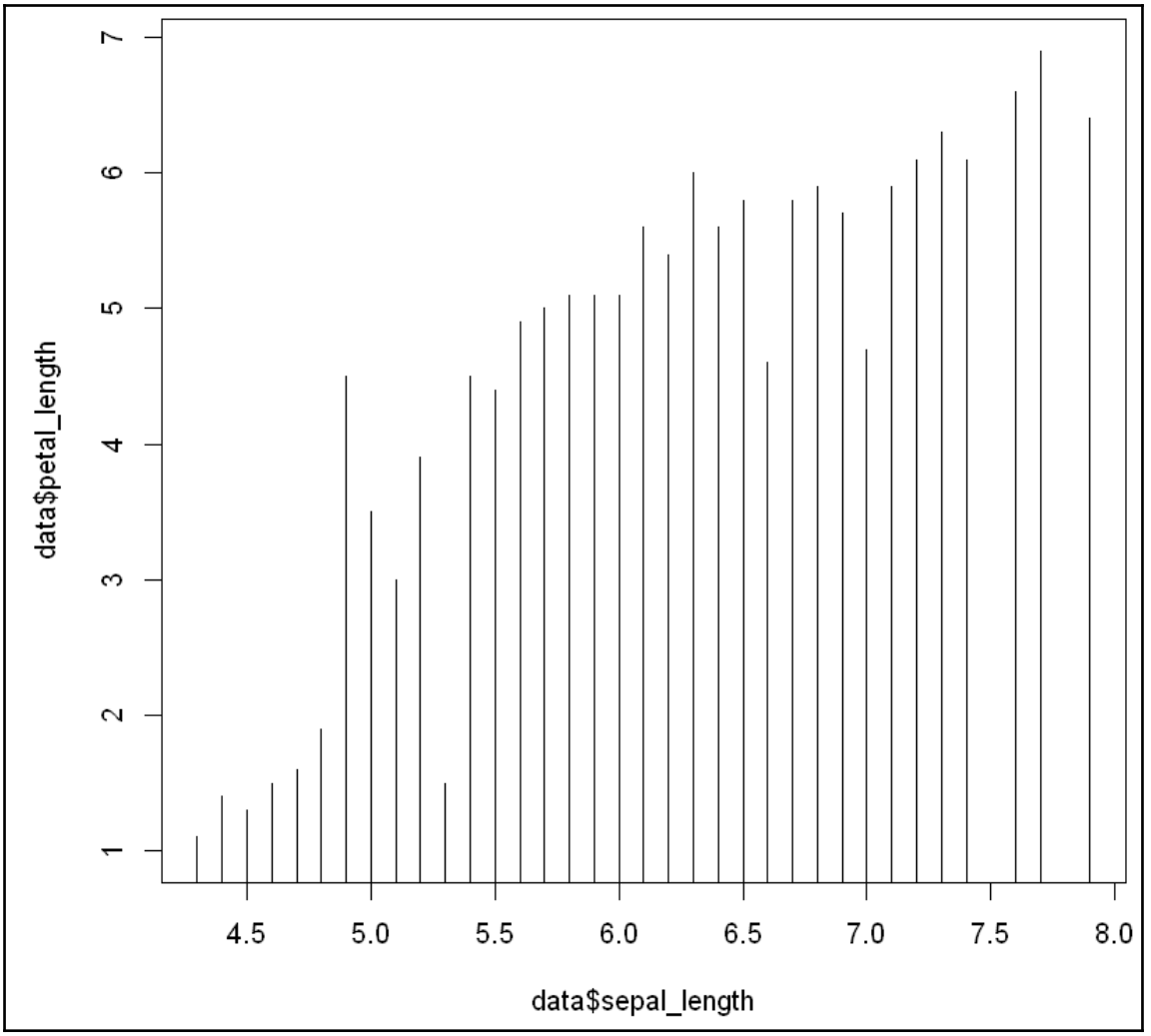

8
7
6
5
4
3
Cylinders
45
40
35
30
25
20
15
10
MPG
2000
2500
3000
3500
4000
4500
5000
Weight

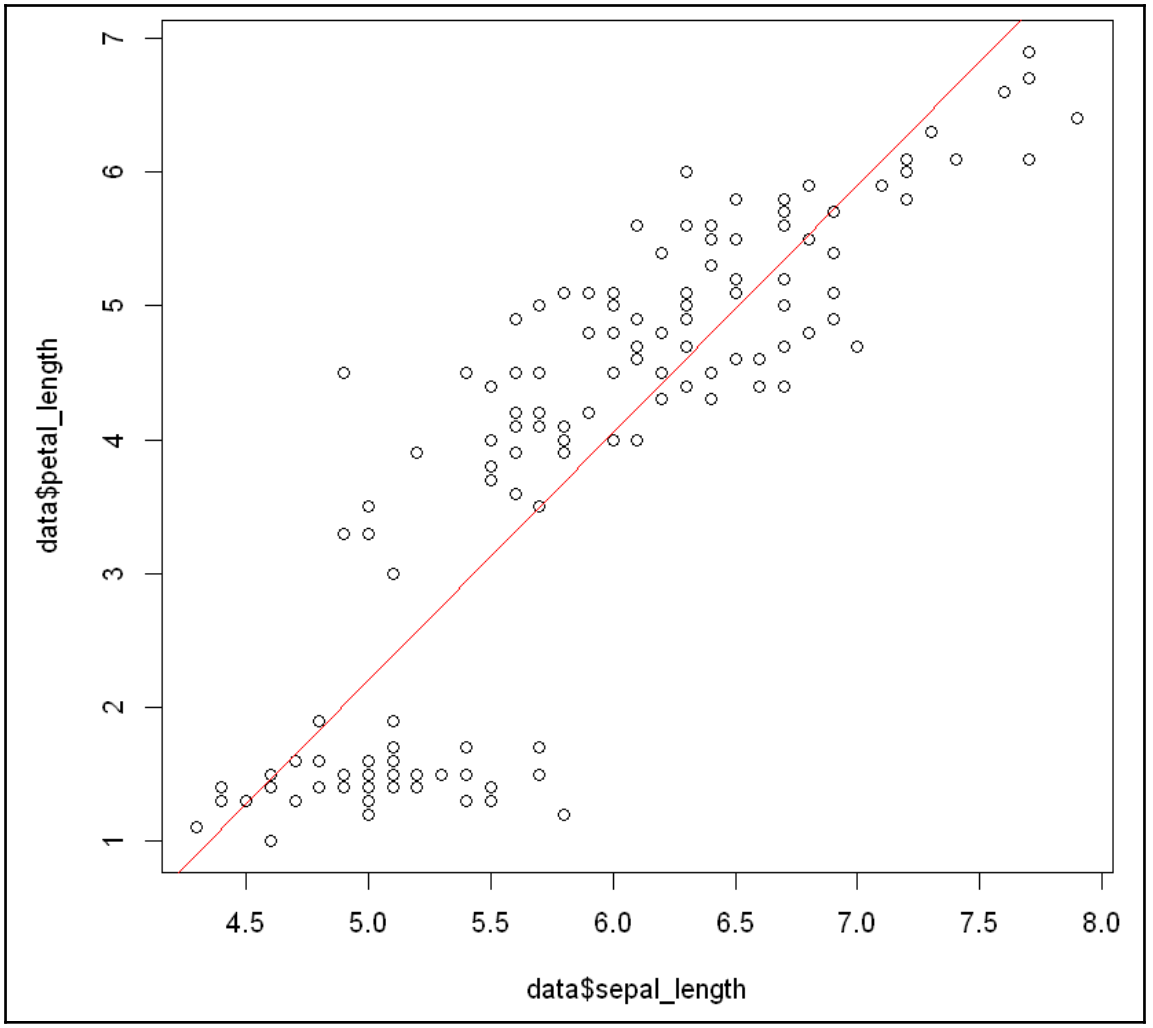

How Many Votes Made by Users
800
600
400
200
0
0
1000
2000
3000
4000
5000
6000
Foo
title
A Text Box

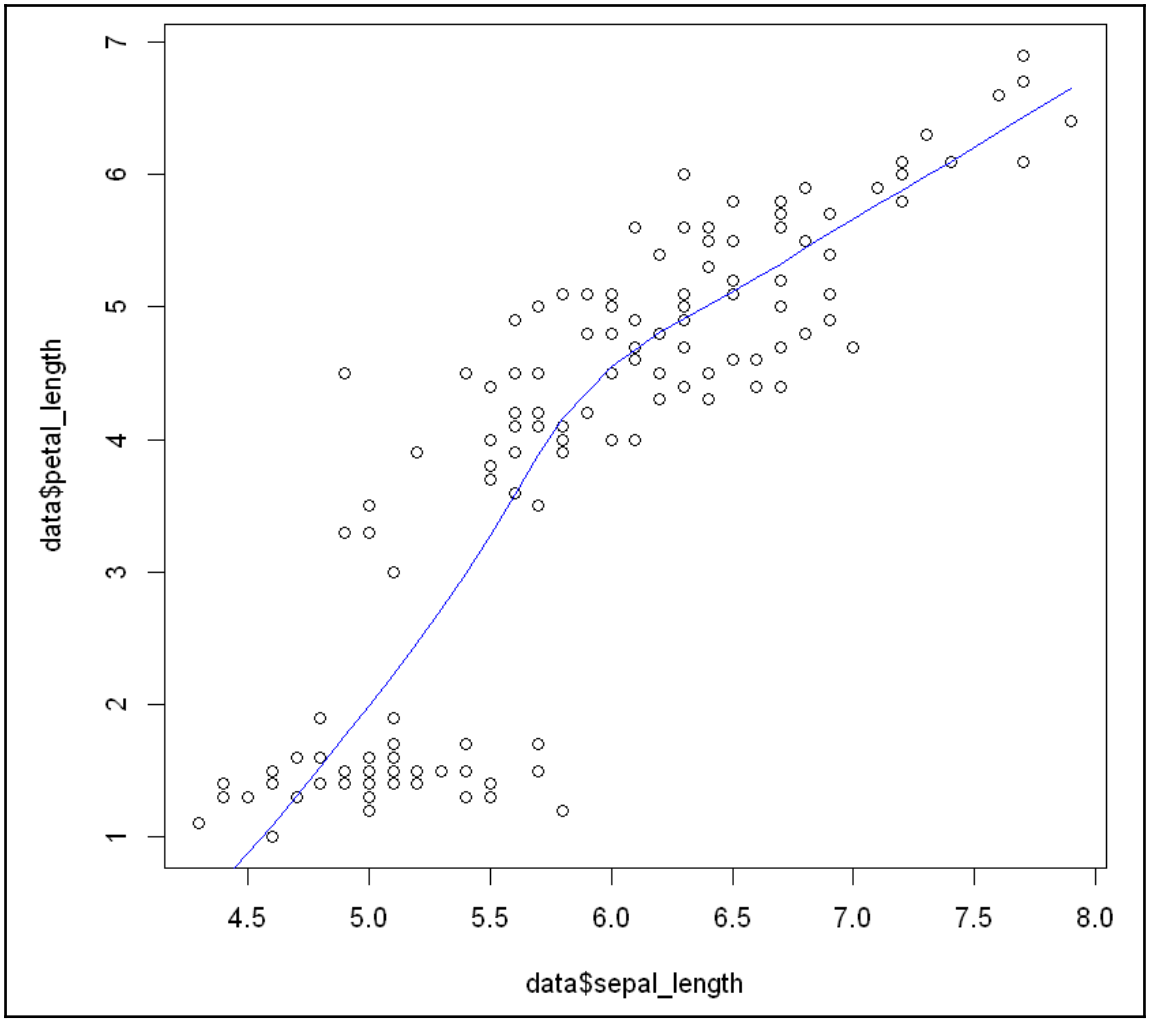
```
 sepal_length    sepal_width     petal_length    petal_width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.400   Median :1.300
Mean   :5.848   Mean   :3.051   Mean   :3.774   Mean   :1.205
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
          species
Iris-setosa    :49
Iris-versicolor:50
Iris-virginica :50
```

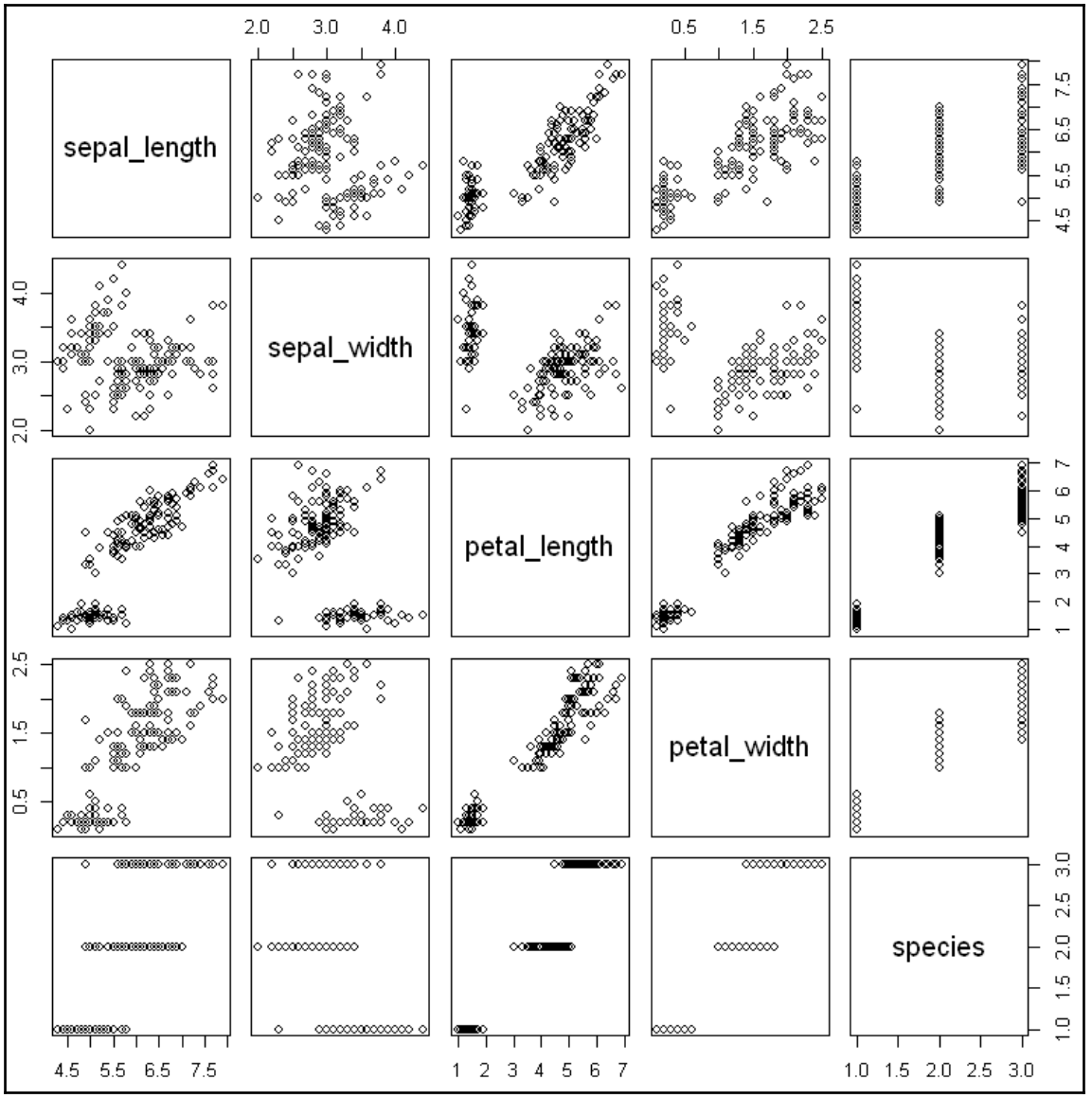

data$petal_length
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
data$sepal_length


data$petal_length
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
data$sepal_length


data$petal_length
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
data$sepal_length

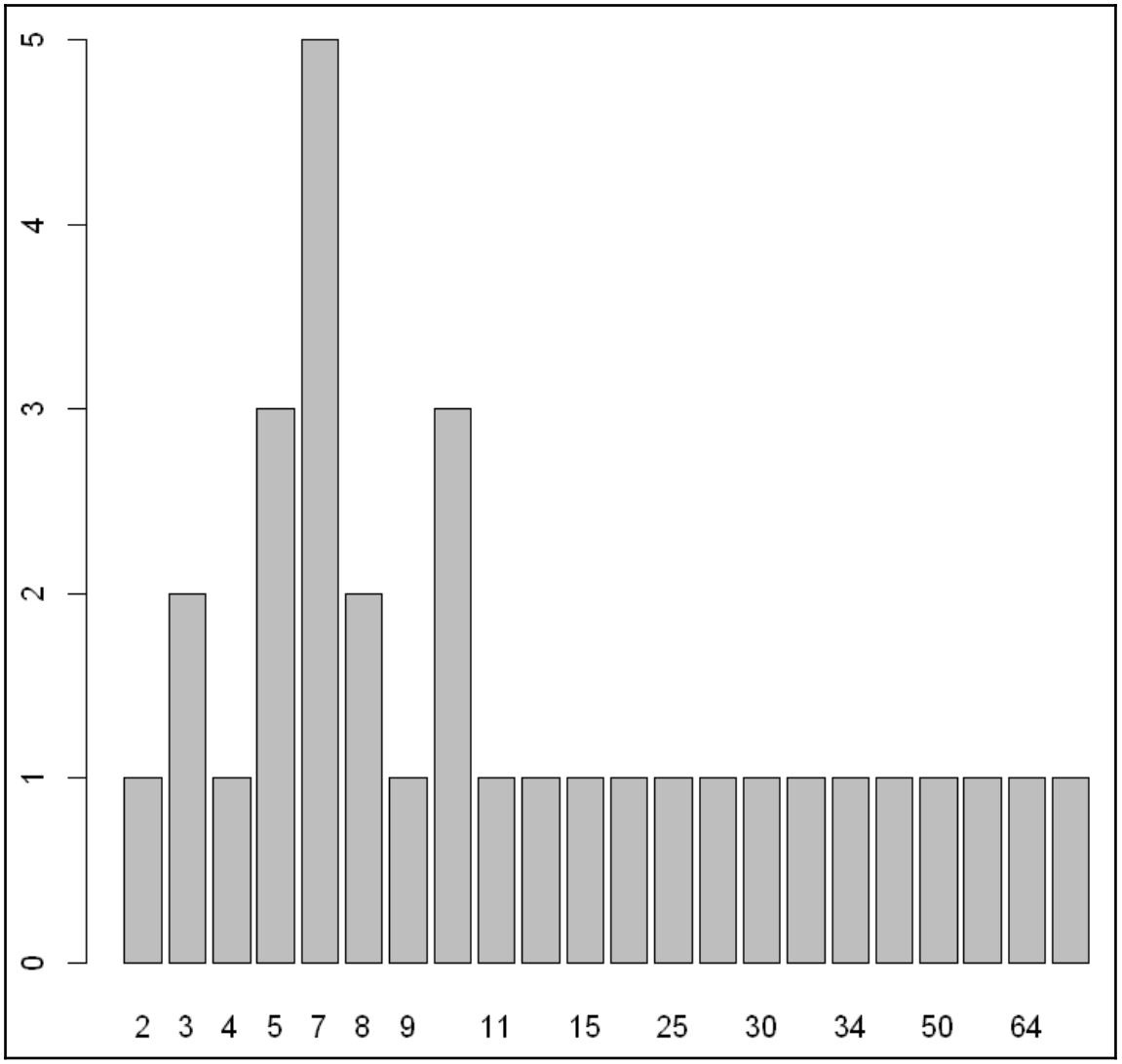

data$petal_length
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
data$sepal_length

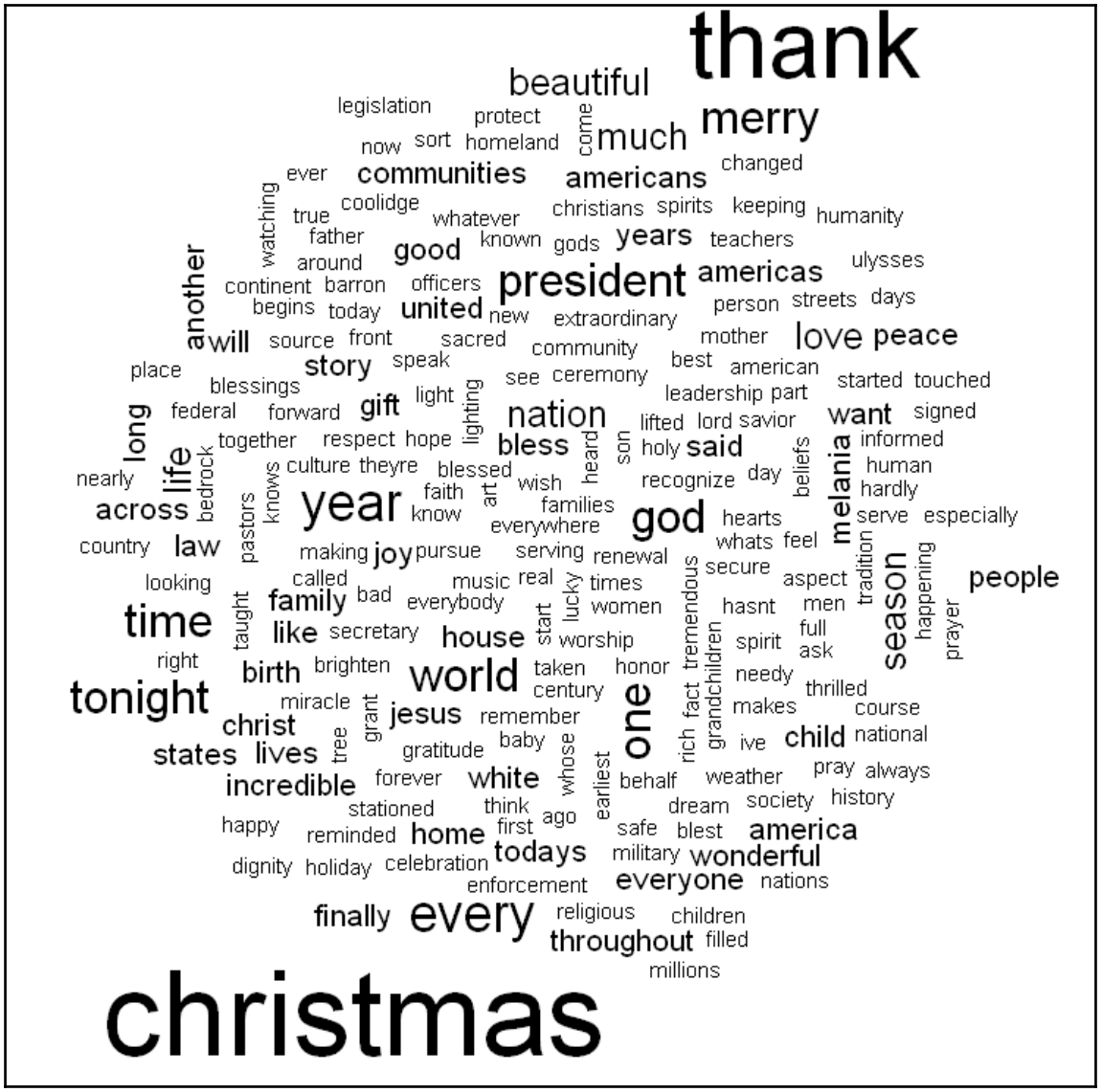

data$petal_length
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
data$sepal_length




```
Number of cases in table: 592
Number of factors: 3
Test for independence of all factors:
        Chisq = 164.92, df = 24, p-value = 5.321e-23
        Chi-squared approximation may be incorrect
```

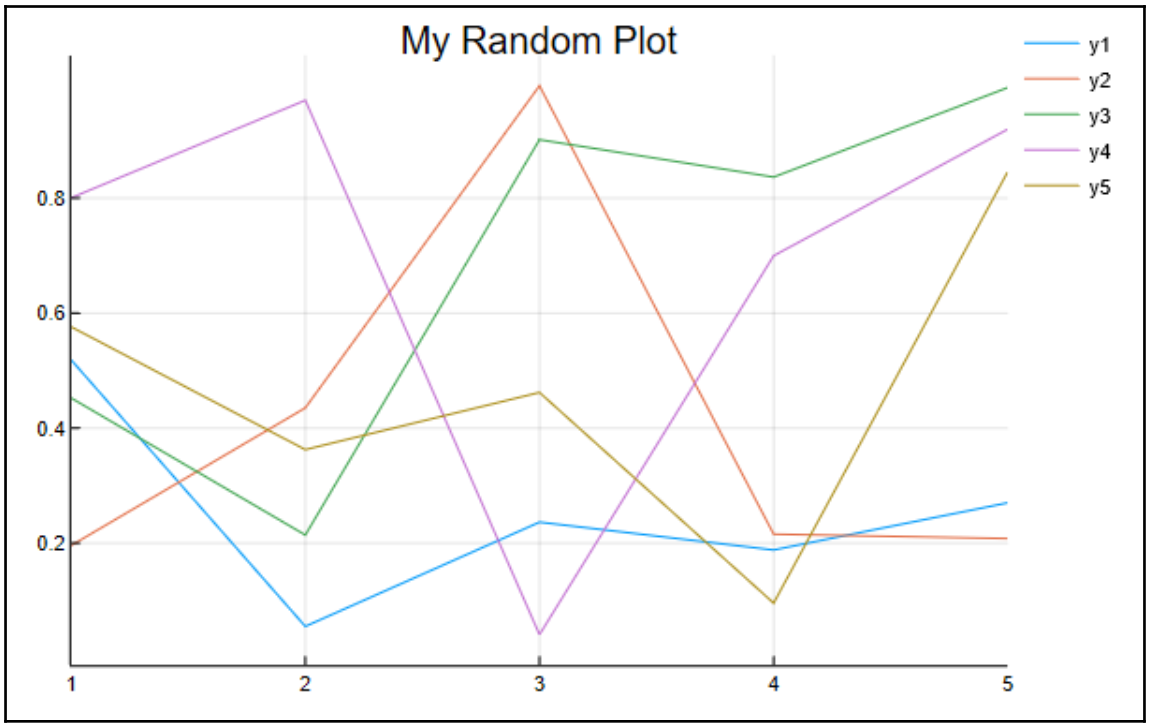
```
, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

, , Sex = Female

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

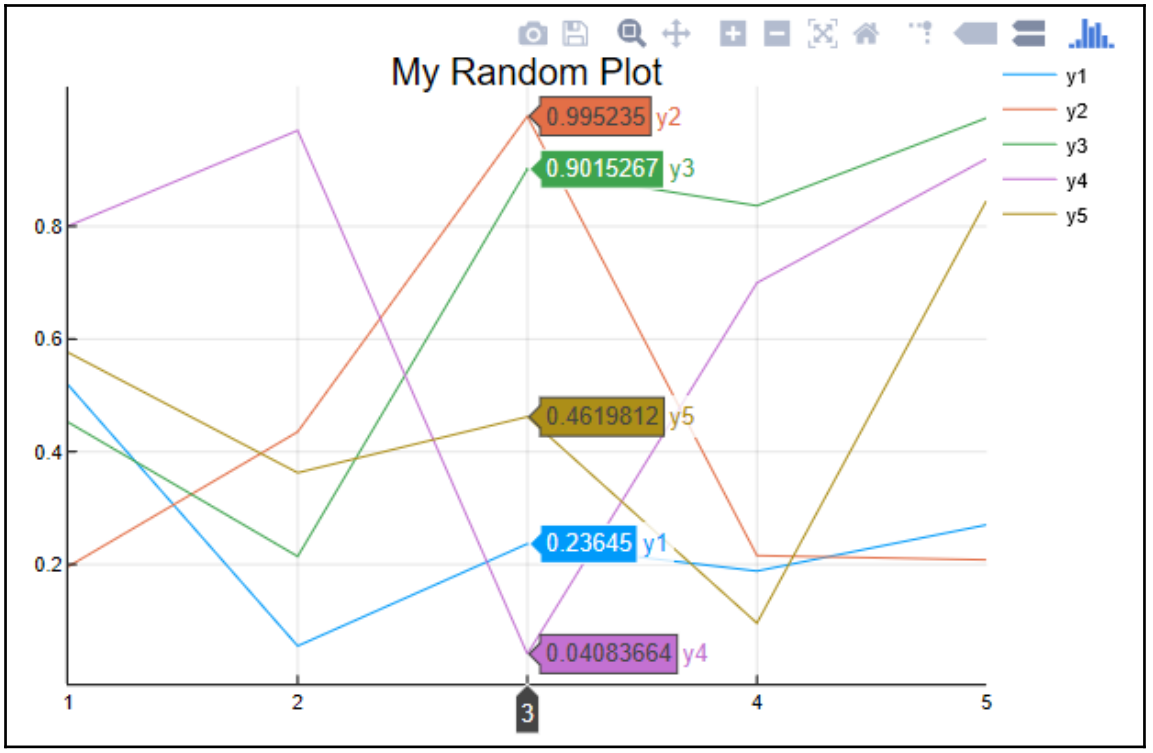

0
1
2
3
4
5
2
3
4
5
7
8
9
11
15
25
30
34
50
64




```
Pkg.update()
```

```
INFO: Cloning cache of AxisAlgorithms from https://github.com/timholy/AxisAlgorithms.jl.git
INFO: Cloning cache of Calculus from https://github.com/johnmyleswhite/Calculus.jl.git
INFO: Cloning cache of CommonSubexpressions from https://github.com/rdeits/CommonSubexpressions.jl.git
INFO: Cloning cache of Compose from https://github.com/GiovineItalia/Compose.jl.git
INFO: Cloning cache of CoupledFields from https://github.com/Mattriks/CoupledFields.jl.git
INFO: Cloning cache of DiffResults from https://github.com/JuliaDiff/DiffResults.jl.git
INFO: Cloning cache of DiffRules from https://github.com/JuliaDiff/DiffRules.jl.git
```

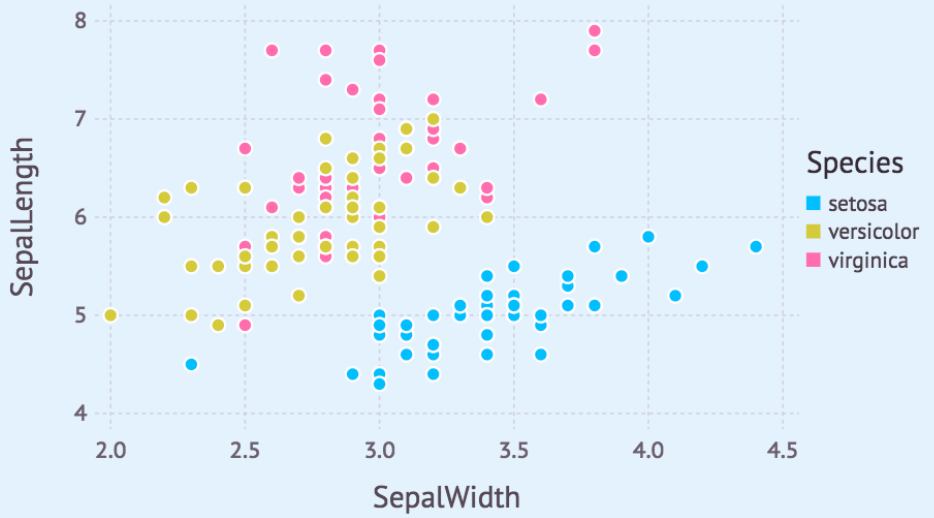

My Random Plot
y1
y2
y3
y4
y5
0.8
0.6
0.4
0.2
1
2
3
4
5




```
In [2]: #Pkg.add("RDatasets")
        #Pkg.add("Dataframe")
        #Pkg.add("Gadfly")
        using RDatasets, DataFrames, Gadfly
        set_default_plot_size(5inch, 5inch/golden)
        describe(dataset("datasets","iris"))
        plot(dataset("datasets","iris"), x="SepalWidth", y="SepalLength",color="Species")
```

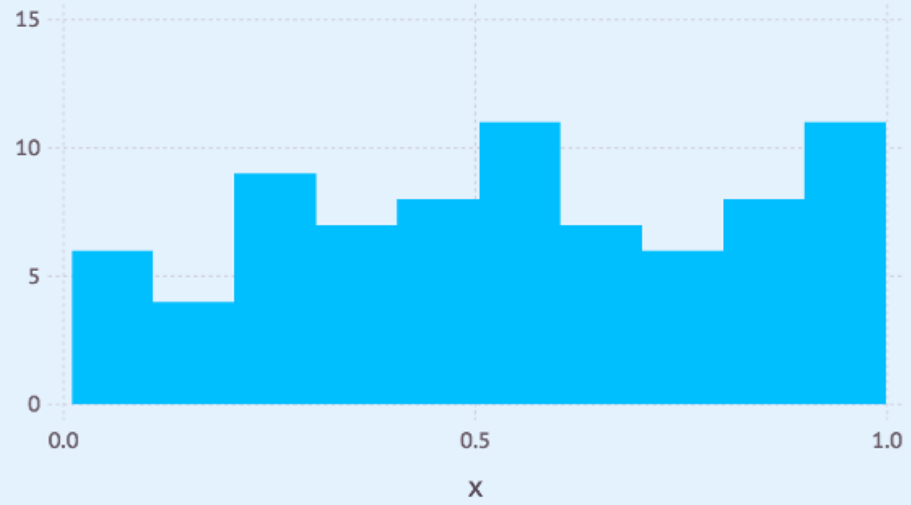
```
SepalLength
Summary Stats:
Mean:           5.843333
Minimum:        4.300000
1st Quartile:   5.100000
Median:         5.800000
3rd Quartile:   6.400000
Maximum:        7.900000
Length:         150
Type:           Float64
Number Missing: 0
% Missing:      0.000000
```

Out[2]:

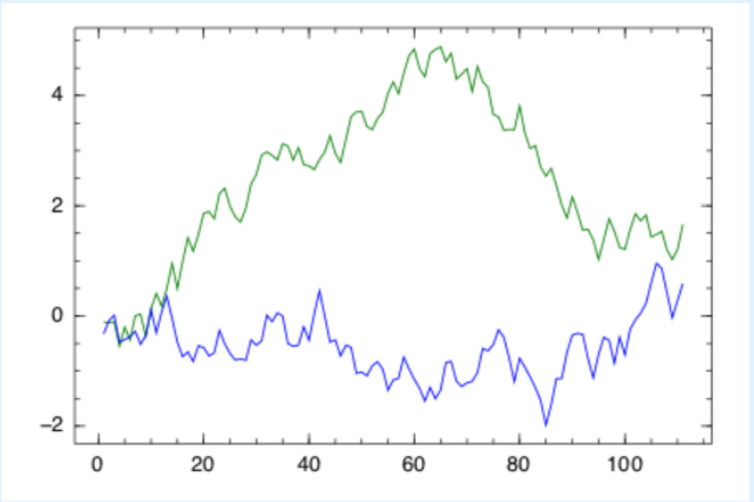

SepalLength
8
7
6
5
4
2.0
2.5
3.0
3.5
4.0
4.5
SepalWidth
Species
setosa
versicolor
virginica

In [3]:
```
using Gadfly
srand(111)
plot(x=rand(77), Geom.histogram(bincount=10))
```

Out[3]:



In [1]:
```
using Winston
srand(111)
pl = plot(cumsum(rand(111) .- 0.5), "g", cumsum(rand(111) .- 0.5), "b")
display(pl)
```



Chapter 5: Working with Widgets

```
In [16]: import ipywidgets as widgets
         from IPython.display import display

In [17]: my_button = widgets.Button(description='Click My Button')
         display(my_button)
```

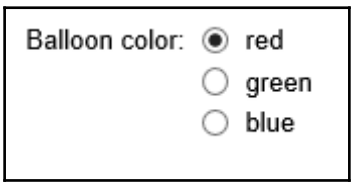
Click My Button

```
In [18]: def my_button_clicked(b):
             print("You clicked on My Button")

In [19]: my_button.on_click(my_button_clicked)
```

```
In [20]: import ipywidgets as widgets
         from IPython.display import display

In [21]: my_button = widgets.Button(description='Click My Button')
         display(my_button)
```

Click My Button

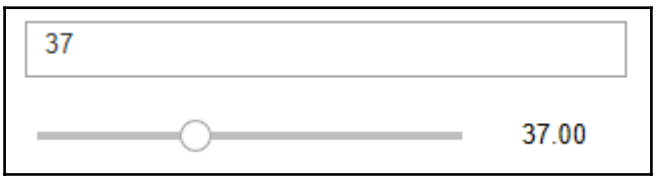
```
You clicked on My Button

In [22]: def my_button_clicked(b):
             print("You clicked on My Button")

In [23]: my_button.on_click(my_button_clicked)
```

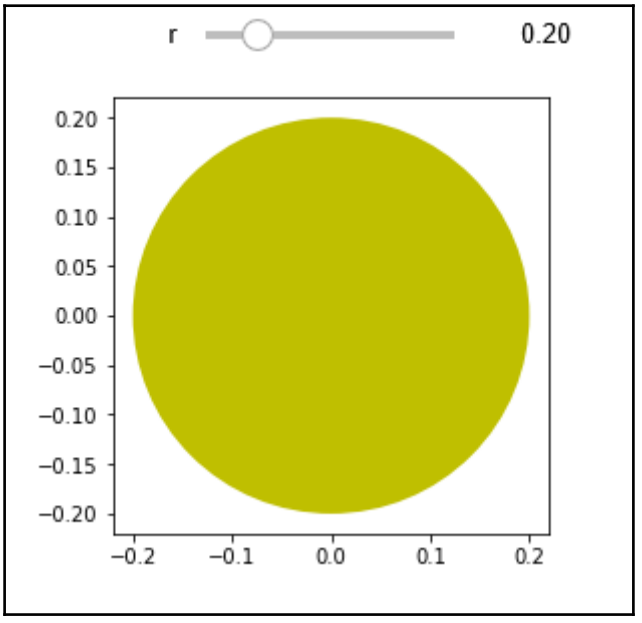


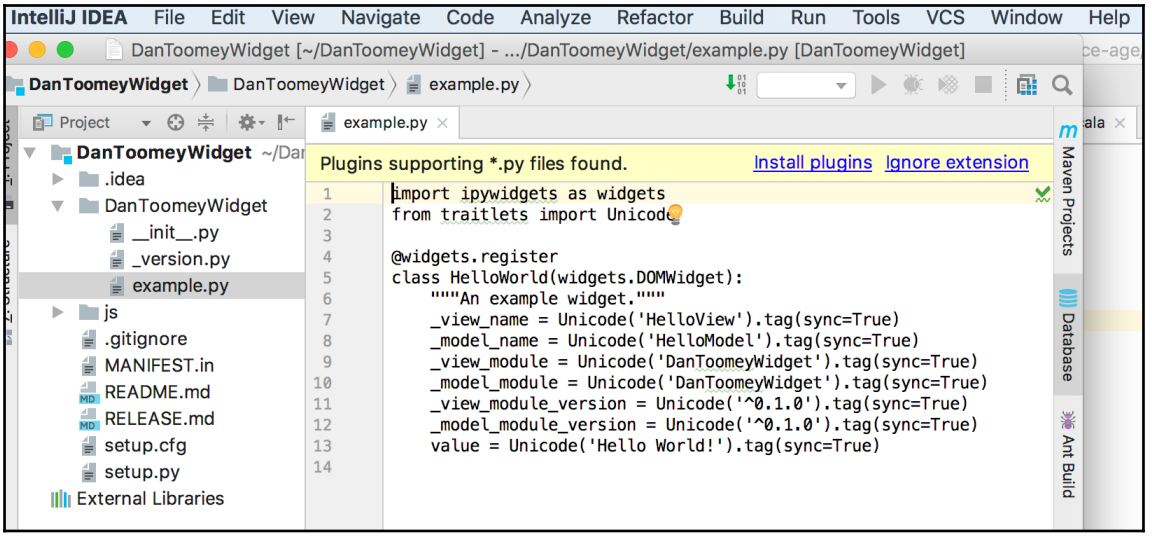
0.00
Hello World

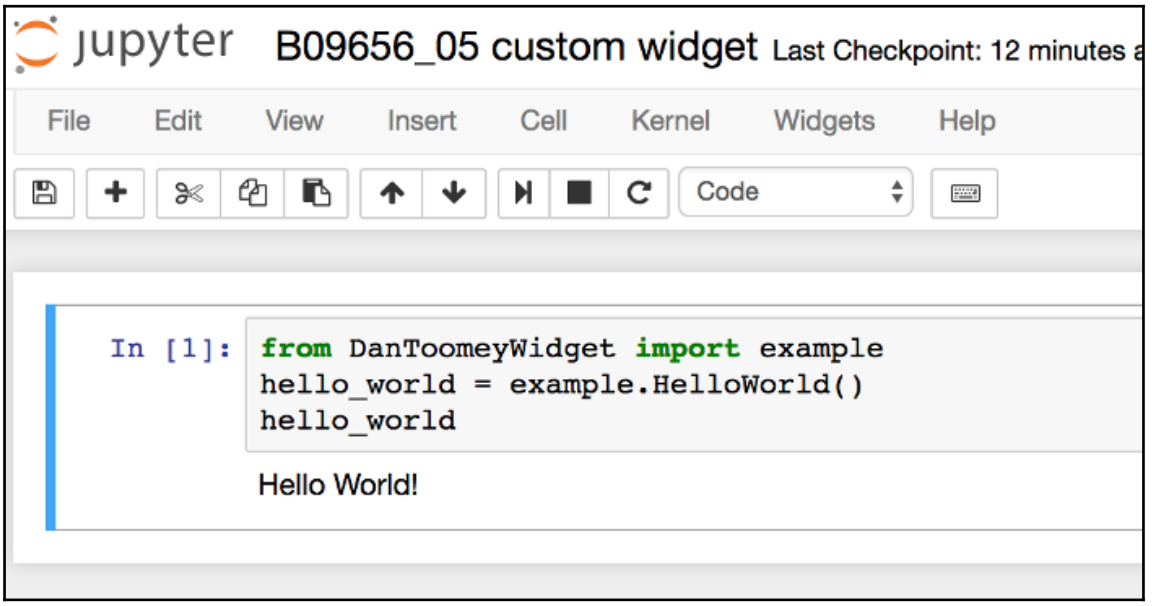
x
True

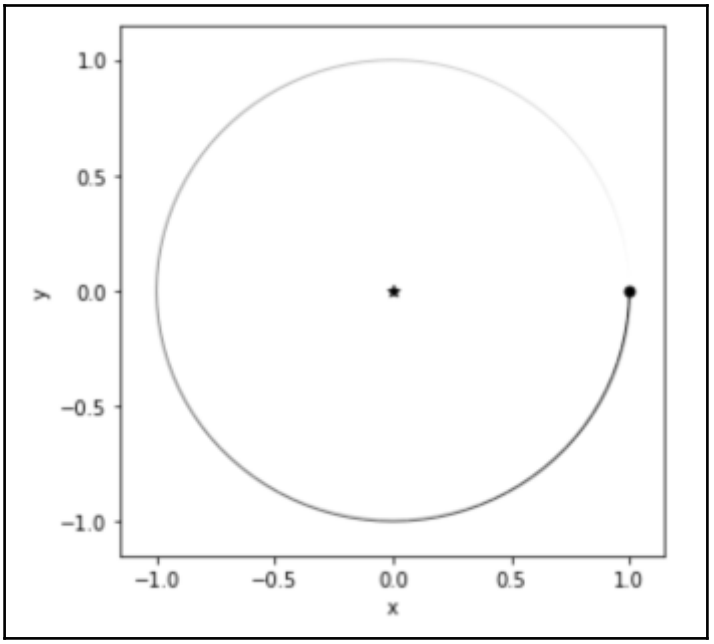
x
Hello Dan
Hello Dan

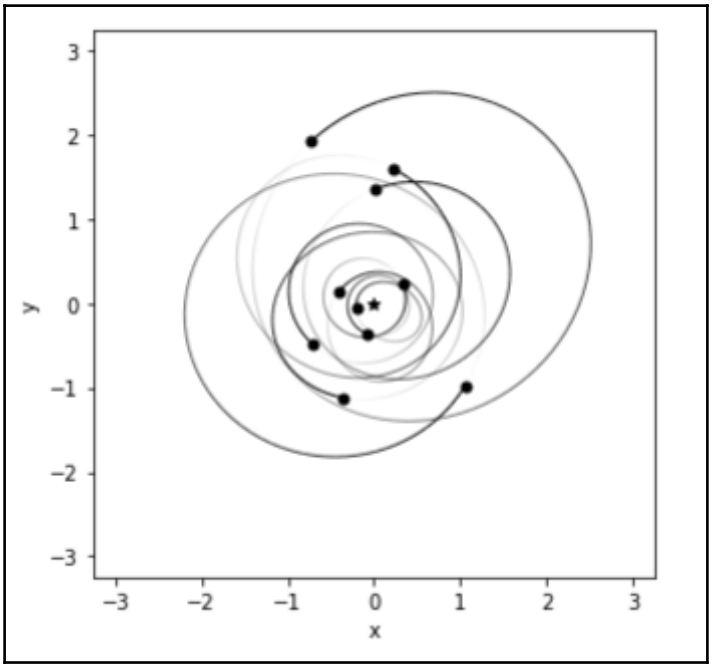

37
37.00


r
0.20
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.2
−0.1
0.0
0.1
0.2

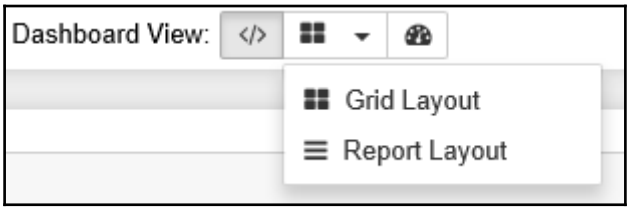

IntelliJ IDEA File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
DanToomeyWidget [~/DanToomeyWidget] - .../DanToomeyWidget/example.py [DanToomeyWidget]
DanToomeyWidget DanToomeyWidget example.py
Project
DanToomeyWidget ~/Dar
.idea
DanToomeyWidget
__init__.py
_version.py
example.py
js
.gitignore
MANIFEST.in
README.md
RELEASE.md
setup.cfg
setup.py
External Libraries
example.py
Plugins supporting *.py files found.
Install plugins
Ignore extension
import ipywidgets as widgets
from traitlets import Unicode
@widgets.register
class HelloWorld(widgets.DOMWidget):
"""An example widget."""
_view_name = Unicode('HelloView').tag(sync=True)
_model_name = Unicode('HelloModel').tag(sync=True)
_view_module = Unicode('DanToomeyWidget').tag(sync=True)
_model_module = Unicode('DanToomeyWidget').tag(sync=True)
_view_module_version = Unicode('^0.1.0').tag(sync=True)
_model_module_version = Unicode('^0.1.0').tag(sync=True)
value = Unicode('Hello World!').tag(sync=True)
Maven Projects
Database
Ant Build

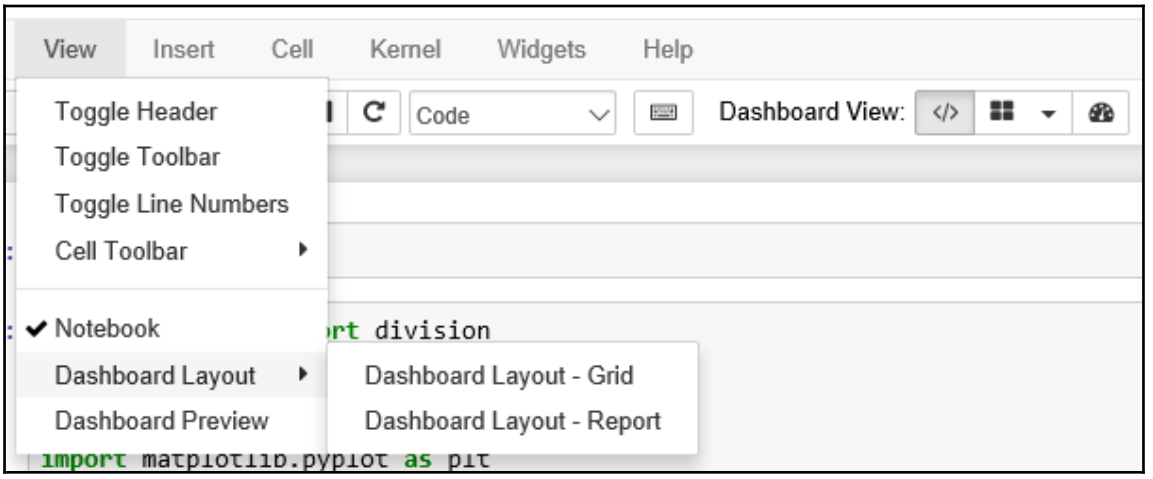

jupyter B09656_05 custom widget Last Checkpoint: 12 minutes a
File Edit View Insert Cell Kernel Widgets Help
Code
In [1]:
from DanToomeyWidget import example
hello_world = example.HelloWorld()
hello_world
Hello World!

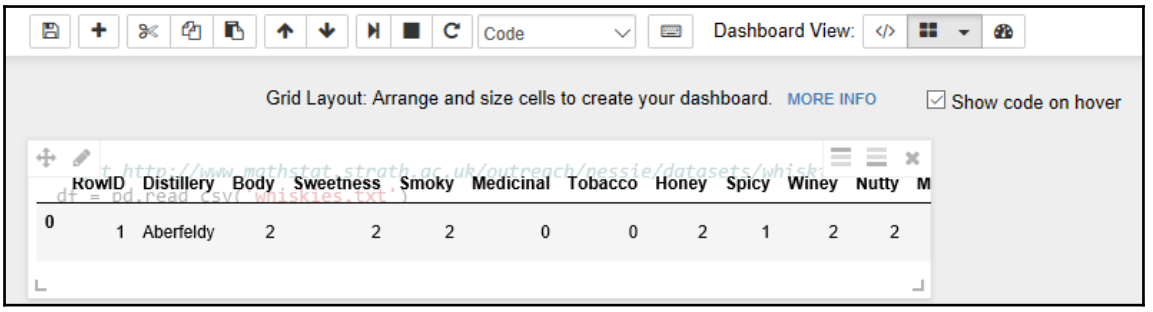

1.0
0.5
0.0
−0.5
−1.0
y
−1.0
−0.5
0.0
0.5
1.0
x


3
2
1
0
−1
−2
−3
y
−3
−2
−1
0
1
2
3
x

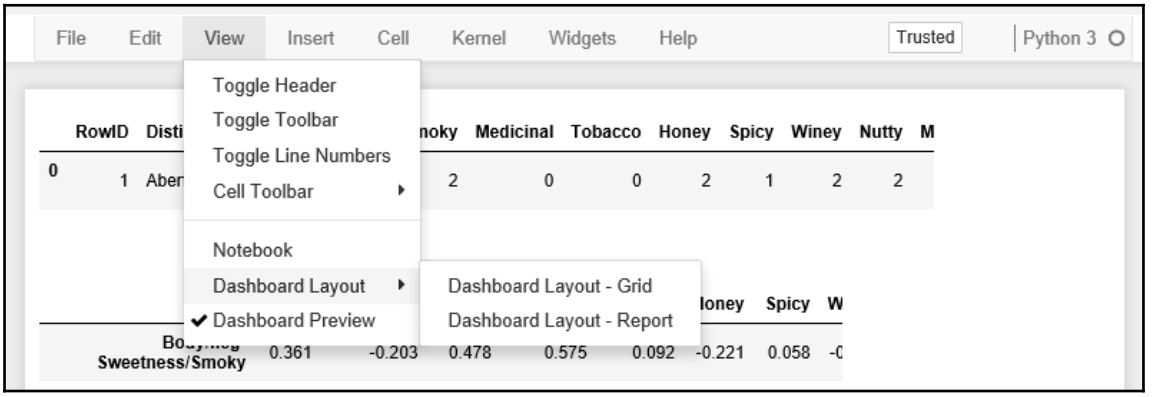
Chapter 6: Jupyter Dashboards



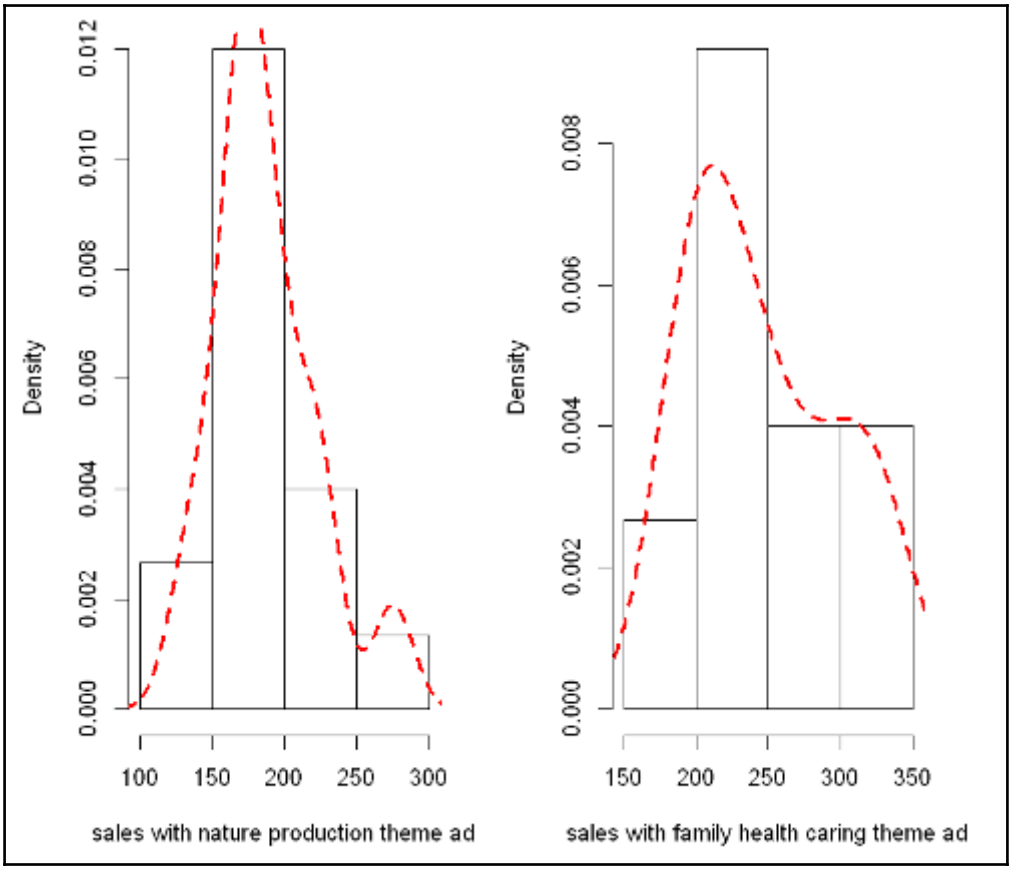


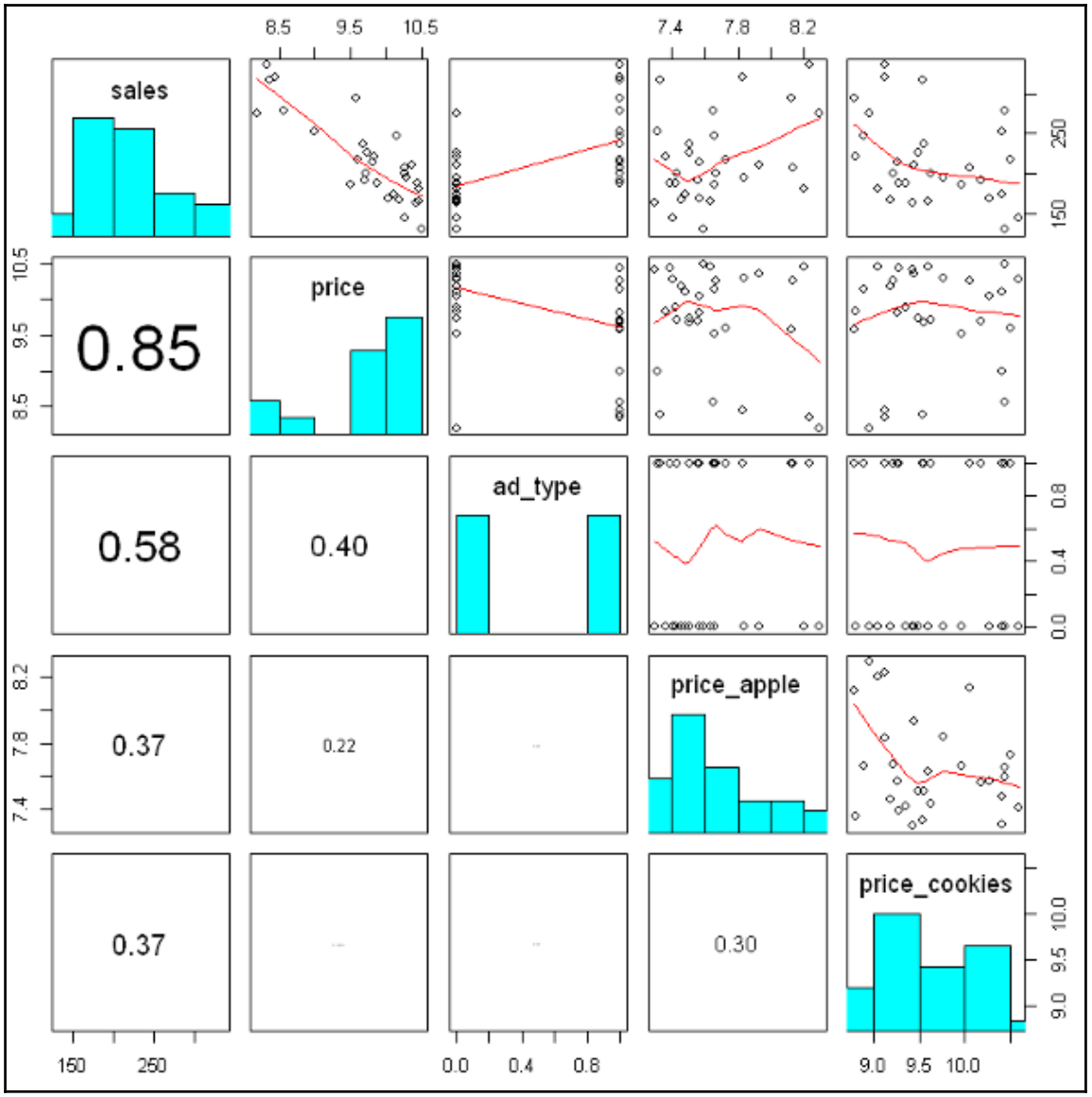


X.cluster.value_counts()
Hidden Cells
matplotlib inline
from __future__ import division
import math


File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Toggle Header
Toggle Toolbar
Toggle Line Numbers
Cell Toolbar
Notebook
Dashboard Layout
Dashboard Preview
Dashboard Layout - Grid
Dashboard Layout - Report
RowID
Medicinal
Tobacco
Honey
Spicy
Winey
Nutty
Sweetness/Smoky
0.361
-0.203
0.478
0.575
0.092
-0.221
0.058

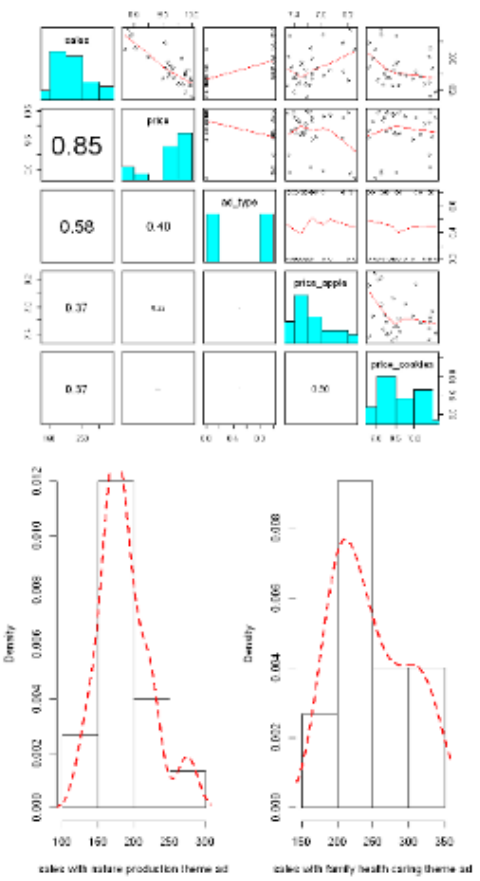
sales	price	ad_type	price_apple	price_cookies
222	9.83	0	7.36	8.80
201	9.72	1	7.43	9.62
247	10.15	1	7.66	8.90
169	10.04	0	7.57	10.26
317	8.38	1	7.33	9.54
227	9.74	0	7.51	9.49




sales
price
ad_type
price_apple
price_cookies
0.85
0.58
0.40
0.37
0.22
0.37
0.30
8.5
9.5
10.5
7.4
7.8
8.2
150
250
0.0
0.4
0.8
9.0
9.5
10.0

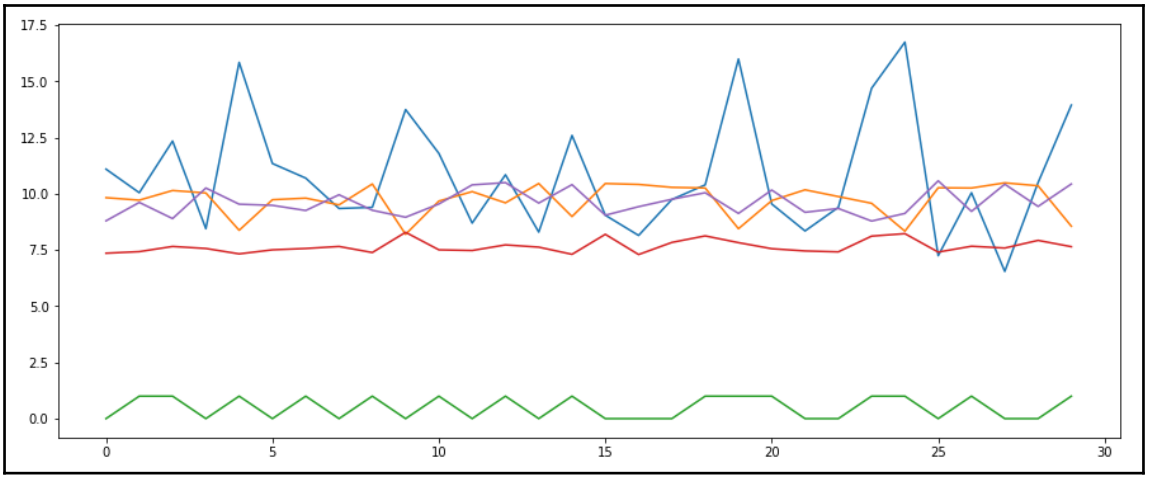
```
Call:
lm(formula = sales ~ price + ad_type + price_apple + price_cookies,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-36.290 -10.488   0.884  10.483  29.471

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    774.813    145.349   5.331 1.59e-05 ***
price          -51.239      5.321  -9.630 6.83e-10 ***
ad_type         29.742      7.249   4.103 0.000380 ***
price_apple     22.089     12.512   1.765 0.089710 .
price_cookies  -25.277      6.296  -4.015 0.000477 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.2 on 25 degrees of freedom
Multiple R-squared:  0.8974,    Adjusted R-squared:  0.881
F-statistic: 54.67 on 4 and 25 DF,  p-value: 5.318e-12
```

sales	price	ad_type	price_apple	price_cookie
222	9.83	0	7.36	8.8
201	9.72	1	7.43	9.6
247	10.15	1	7.66	8.9
169	10.04	0	7.57	10.2
317	8.38	1	7.33	9.5
227	9.74	0	7.51	9.4




```
Call:
lm(formula = sales ~ price + ad_type + price_apple + price
_cookies,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-36.290 -10.488   0.884  10.483  29.471

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   774.813    145.349   5.331 1.59e-05 ***
price         -51.239      5.321  -9.630 6.83e-10 ***
ad_type        29.742      7.249   4.103 0.000380 ***
price_apple    22.089     12.512   1.765 0.089710 .
price_cookies -25.277      6.296  -4.015 0.000477 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 18.2 on 25 degrees of freedom
Multiple R-squared:  0.8974,    Adjusted R-squared:  0.88
1
F-statistic: 54.67 on 4 and 25 DF,  p-value: 5.318e-12
```

	sales	price	ad_type	price_apple	price_cookies
0	222	9.83	0	7.36	8.80
1	201	9.72	1	7.43	9.62
2	247	10.15	1	7.66	8.90
3	169	10.04	0	7.57	10.26
4	317	8.38	1	7.33	9.54


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
5
10
15
20
25
30

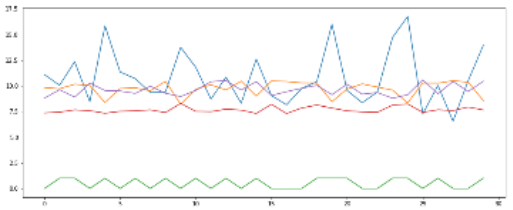
OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.987
Model:	OLS	**Adj. R-squared:**	0.985
Method:	Least Squares	**F-statistic:**	493.2
Date:	Mon, 18 Dec 2017	**Prob (F-statistic):**	2.20e-23
Time:	22:31:35	**Log-Likelihood:**	-47.244
No. Observations:	29	**AIC:**	102.5
Df Residuals:	25	**BIC:**	108.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
price	-1.8267	0.359	-5.089	0.000	-2.566	-1.087
ad_type	1.8703	0.520	3.599	0.001	0.800	2.941
price_apple	3.9575	0.475	8.339	0.000	2.980	4.935
price_cookies	-0.2758	0.391	-0.705	0.488	-1.082	0.530

Omnibus:	0.695	**Durbin-Watson:**	2.175
Prob(Omnibus):	0.706	**Jarque-Bera (JB):**	0.656
Skew:	0.327	**Prob(JB):**	0.720
Kurtosis:	2.661	**Cond. No.**	37.9

	sales	price	ad_type	price_apple	price_cookies
0	222	9.83	0	7.36	8.80
1	201	9.72	1	7.43	9.62
2	247	10.15	1	7.66	8.90
3	169	10.04	0	7.57	10.26
4	317	8.38	1	7.33	9.54



OLS Regression Results

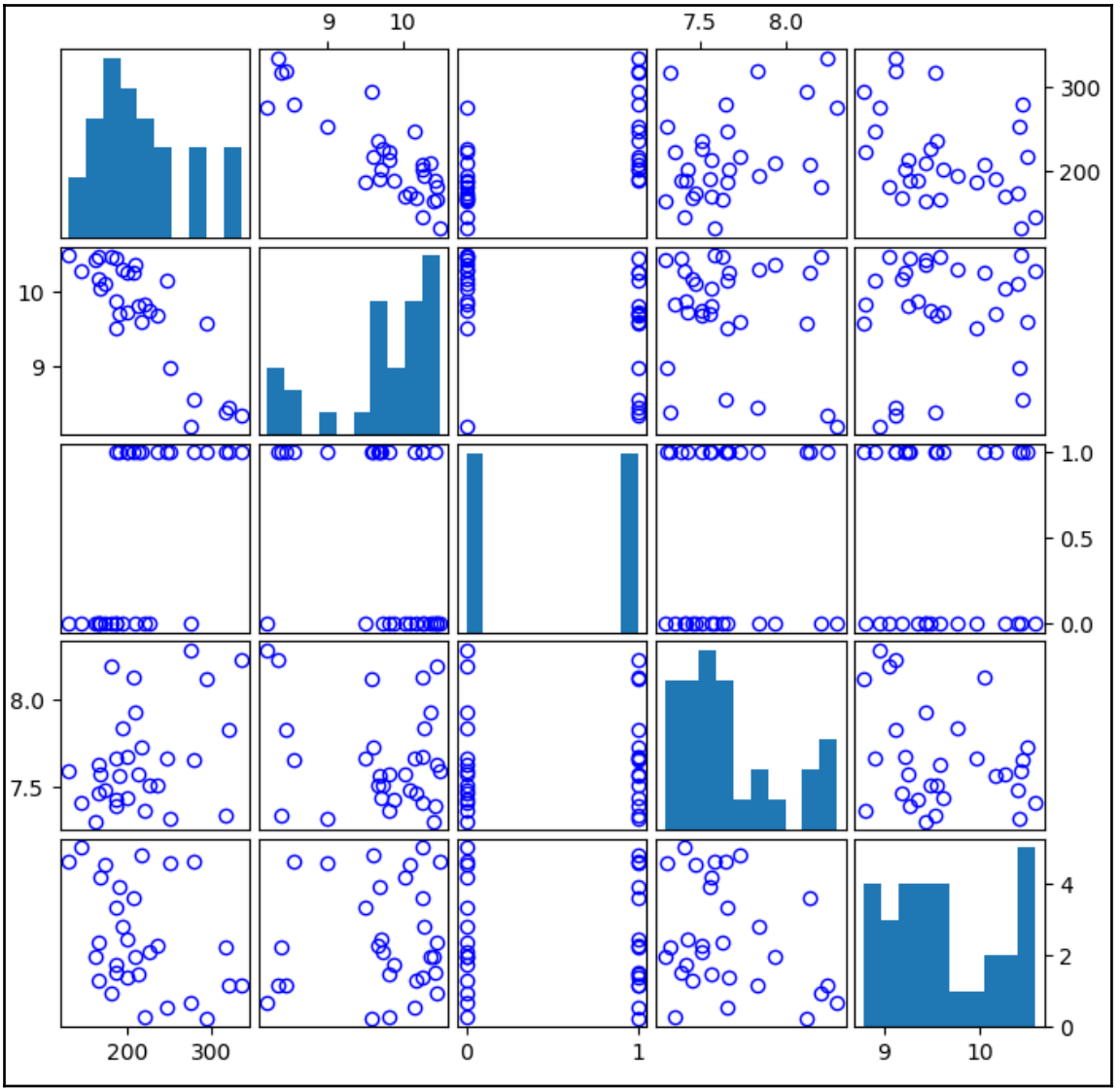
Dep. Variable:	sales	**R-squared:**	0.987
Model:	OLS	**Adj. R-squared:**	0.985
Method:	Least Squares	**F-statistic:**	493.2
Date:	Mon, 18 Dec 2017	**Prob (F-statistic):**	2.20e-23
Time:	22:31:35	**Log-Likelihood:**	-47.244
No. Observations:	29	**AIC:**	102.5
Df Residuals:	25	**BIC:**	108.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
price	-1.8267	0.359	-5.089	0.000	-2.566	-1.087
ad_type	1.8703	0.520	3.599	0.001	0.800	2.941
price_apple	3.9575	0.475	8.339	0.000	2.980	4.935
price_cookies	-0.2758	0.391	-0.705	0.488	-1.082	0.530

Omnibus:	0.695	**Durbin-Watson:**	2.175
Prob(Omnibus):	0.706	**Jarque-Bera (JB):**	0.656
Skew:	0.327	**Prob(JB):**	0.720
Kurtosis:	2.661	**Cond. No.**	37.9

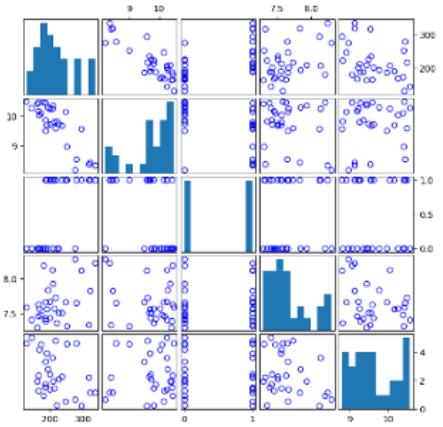
	sales	price	ad_type	price_apple	price_cookies
1	222	9.83	0	7.36	8.8
2	201	9.72	1	7.43	9.62
3	247	10.15	1	7.66	8.9
4	169	10.04	0	7.57	10.26
5	317	8.38	1	7.33	9.54
6	227	9.74	0	7.51	9.49

```
Formula: sales ~ 1 + price + ad_type + price_apple + price_cookies

Coefficients:
               Estimate Std.Error  t value Pr(>|t|)
(Intercept)     774.813   145.349   5.3307    <1e-4
price           -51.2393   5.32094 -9.62976    <1e-9
ad_type          29.7417   7.24851  4.10314   0.0004
price_apple      22.0892   12.5123   1.7654   0.0897
price_cookies   -25.2766   6.29589 -4.01478   0.0005
```


9
10
7.5
8.0
300
200
10
9
1.0
0.5
0.0
8.0
7.5
4
2
0
200
300
0
1
9
10

	sales	price	ad_type	price_apple	price_coo
1	222	9.83	0	7.36	
2	201	9.72	1	7.43	
3	247	10.15	1	7.66	
4	169	10.04	0	7.57	
5	317	8.38	1	7.33	
6	227	9.74	0	7.51	

```
DataFrames.DataFrameRegressionModel{GLM.LinearModel{GLM.LmResp{Array
{Float64,1}},GLM.DensePredChol{Float64,Base.LinAlg.Cholesky{Float64,
Array{Float64,2}}}},Array{Float64,2}}

Formula: sales ~ 1 + price + ad_type + price_apple + price_cookies

Coefficients:
               Estimate Std.Error  t value Pr(>|t|)
(Intercept)     774.813   145.349   5.3307    <1e-4
price           -51.2393  5.32094 -9.62976    <1e-9
ad_type          29.7417  7.24851  4.10314   0.0004
price_apple      22.0892  12.5123   1.7654   0.0897
price_cookies   -25.2766  6.29589 -4.01478   0.0005
```

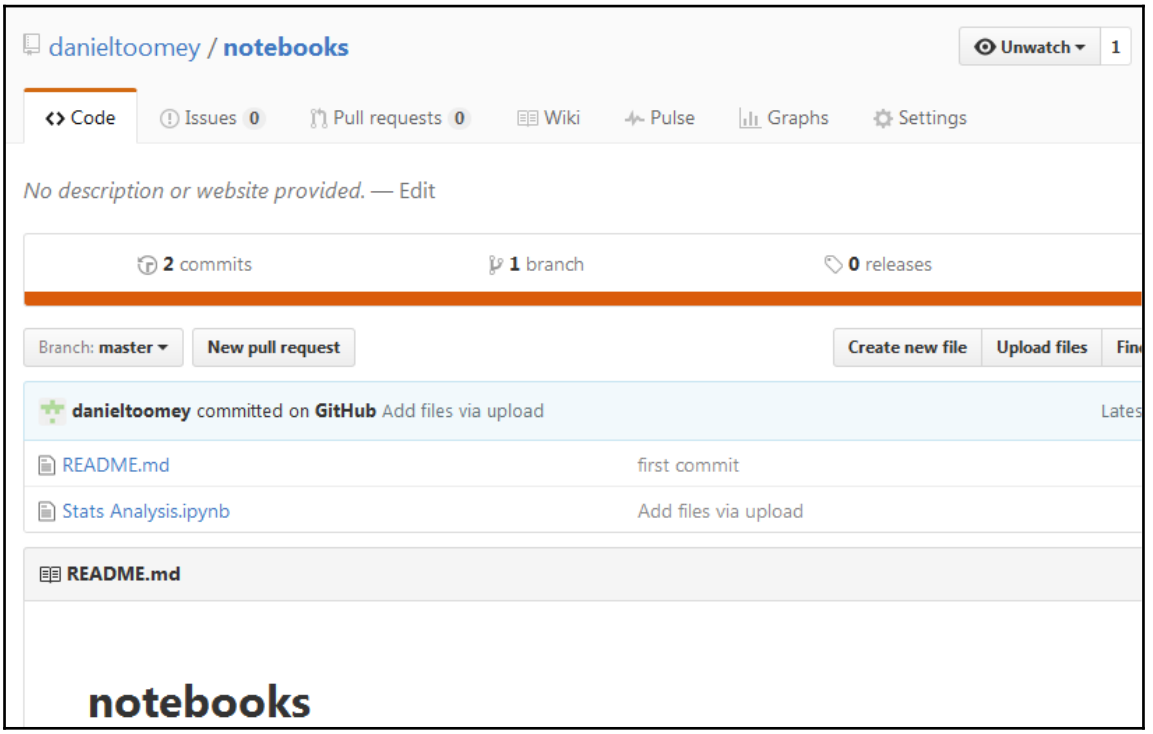


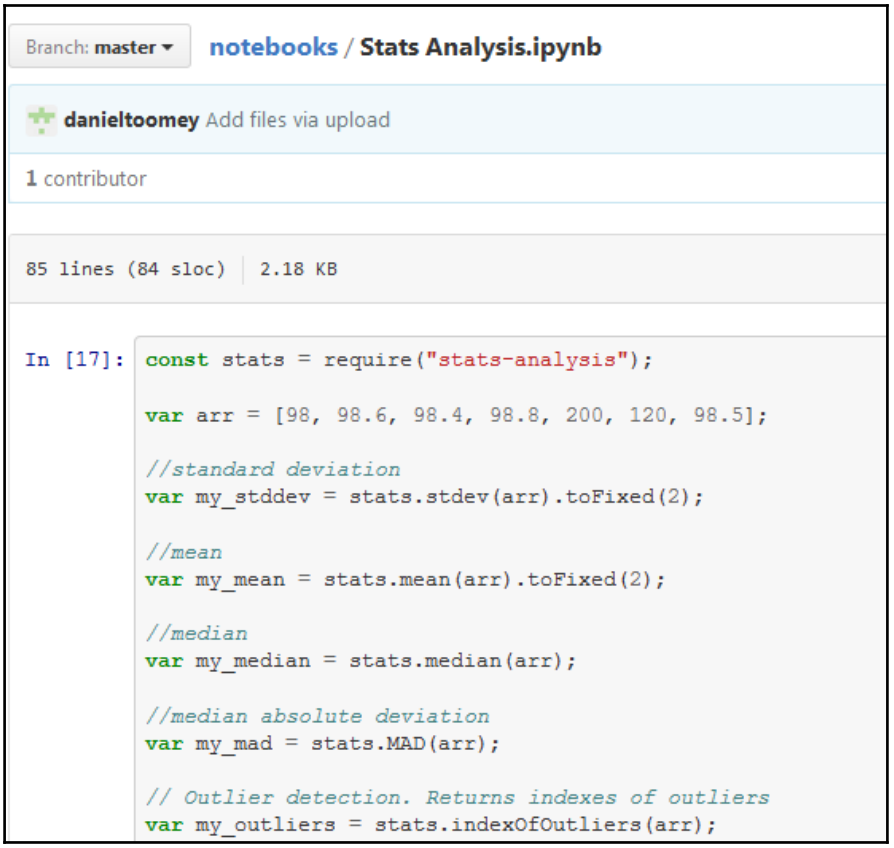
```
[
    {
        "name": "lion",
        "avg_weight": "400"
    },
    {
        "name": "tiger",
        "avg_weight": "400"
    },
    {
        "name": "human",
        "avg_weight": "150"
    },
    {
        "name": "elephant",
        "avg_weight": "2000"
    }
]
2000
```

```
Raw data is  [ 98, 98.6, 98.4, 98.8, 200, 120, 98.5 ]
Standard Deviation is  35.07
Mean is  116.04
Median is  98.6
Median Abs Deviation is 0.20000000000000284
The outliers of the data set are  [ 4, 5, 6 ]
The data set without outliers is  [ 98, 98.6, 98.4, 98.8 ]
```

```
[                                       Raw data is  [ 98, 98.6, 98.4, 98.8, 200, 120, 98.5 ]
    {                                   Standard Deviation is  35.07
        "name": "lion",                 Mean is  116.04
        "avg_weight": "400"             Median is  98.6
    },                                  Median Abs Deviation is 0.20000000000000284
    {                                   The outliers of the data set are  [ 4, 5, 6 ]
        "name": "tiger",                The data set without outliers is  [ 98, 98.6, 98.4, 98.8 ]
        "avg_weight": "400"
    },
    {
        "name": "human",
        "avg_weight": "150"
    },
    {
        "name": "elephant",
        "avg_weight": "2000"
    }
]
2000
```

Chapter 7: Sharing Your Code



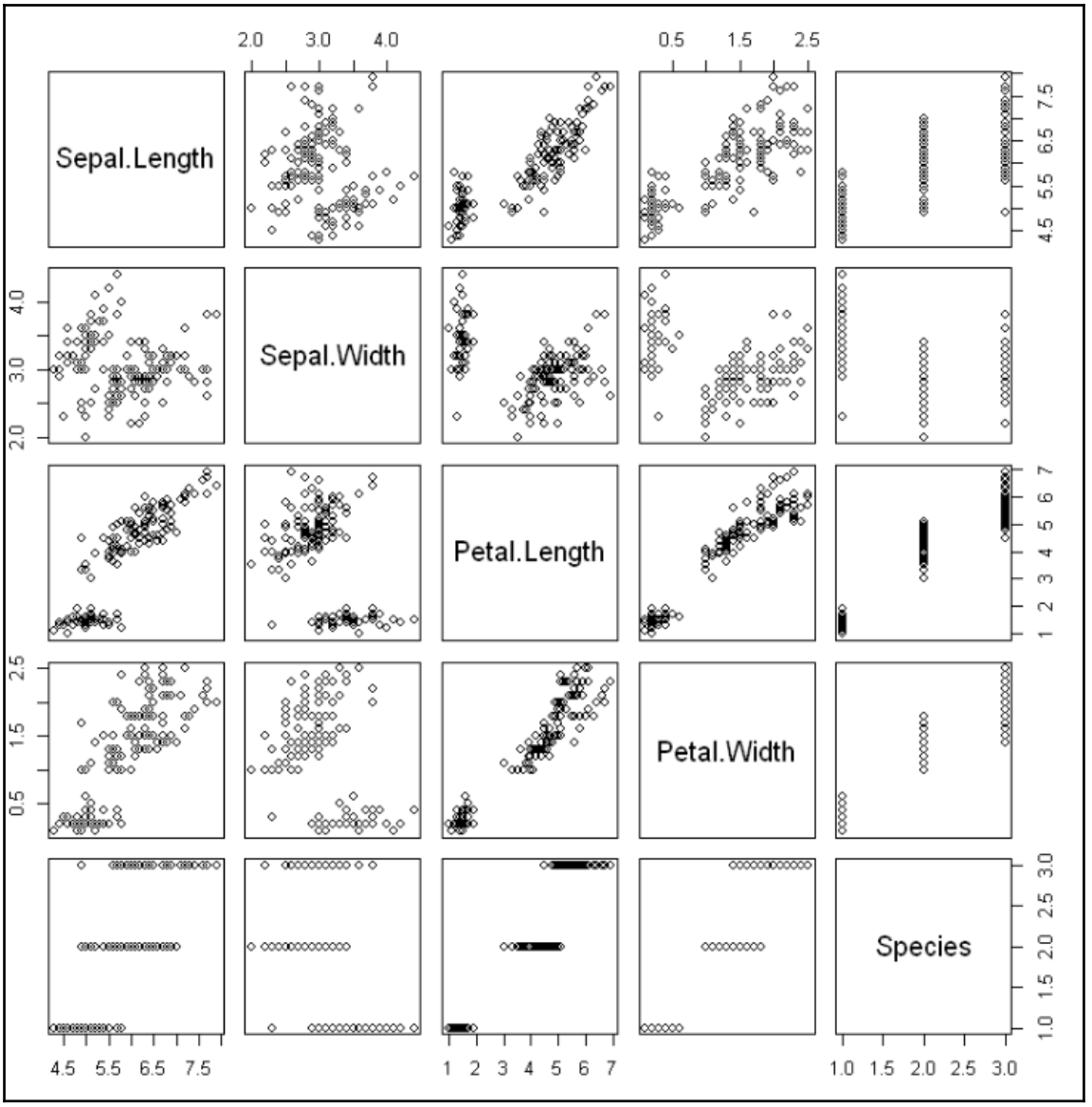

Branch: master
notebooks / Stats Analysis.ipynb
danieltoomey Add files via upload
1 contributor
85 lines (84 sloc) 2.18 KB
In [17]:
const stats = require("stats-analysis");
var arr = [98, 98.6, 98.4, 98.8, 200, 120, 98.5];
//standard deviation
var my_stddev = stats.stdev(arr).toFixed(2);
//mean
var my_mean = stats.mean(arr).toFixed(2);
//median
var my_median = stats.median(arr);
//median absolute deviation
var my_mad = stats.MAD(arr);
// Outlier detection. Returns indexes of outliers
var my_outliers = stats.indexOfOutliers(arr);

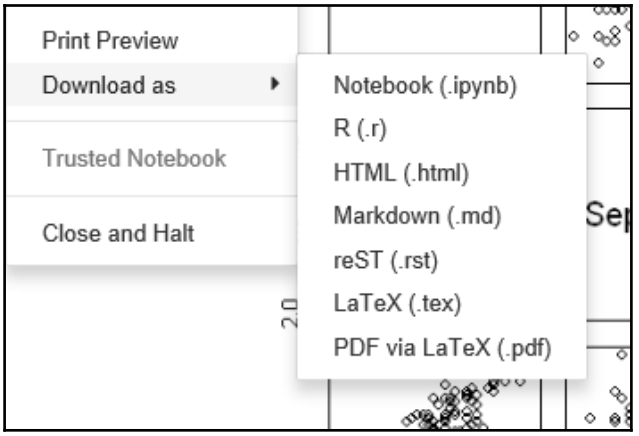
```
data(iris)
head(iris)
```

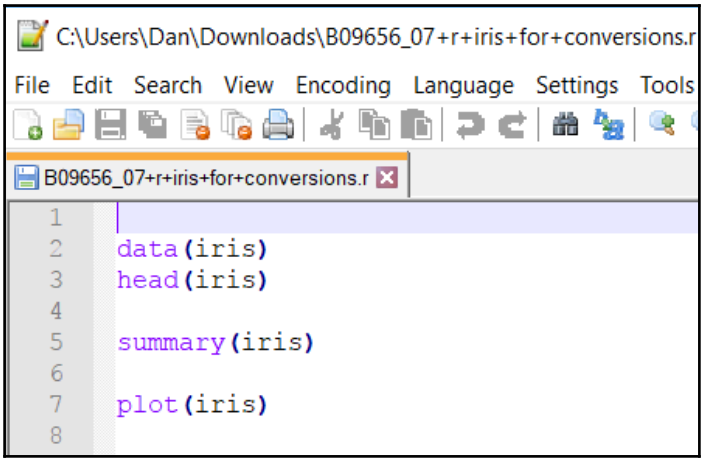
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

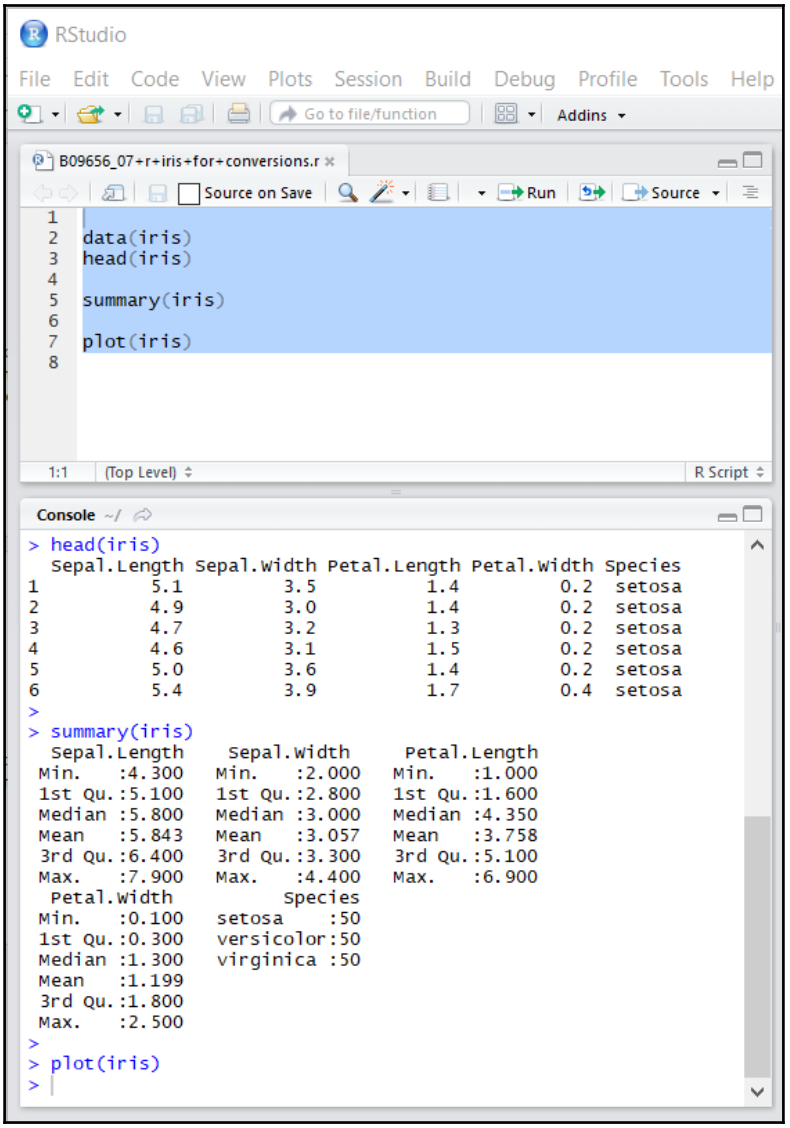
```
summary(iris)
```

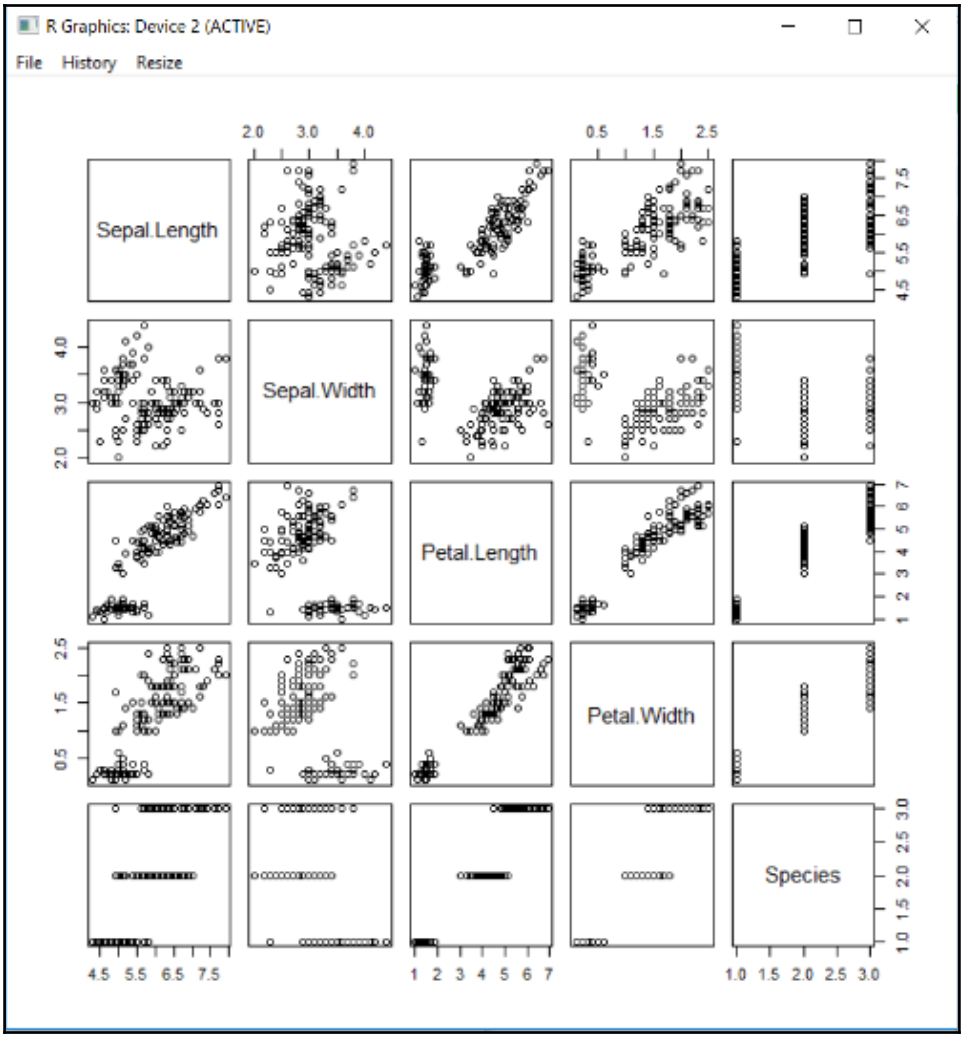
```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

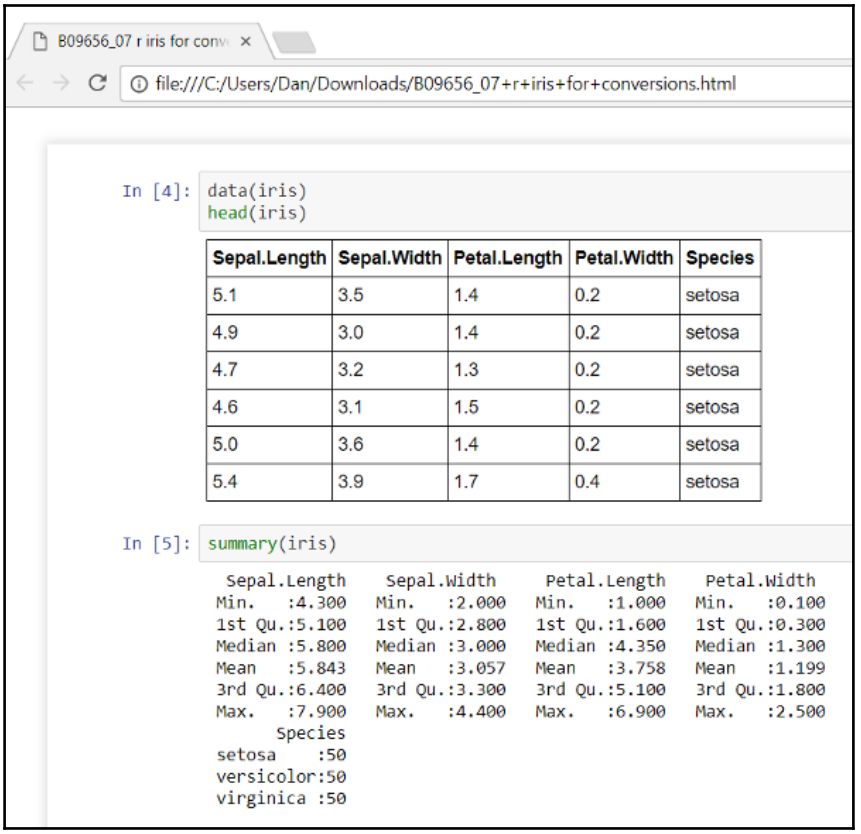

Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species


Print Preview
Download as
Trusted Notebook
Close and Halt
Notebook (.ipynb)
R (.r)
HTML (.html)
Markdown (.md)
reST (.rst)
LaTeX (.tex)
PDF via LaTeX (.pdf)

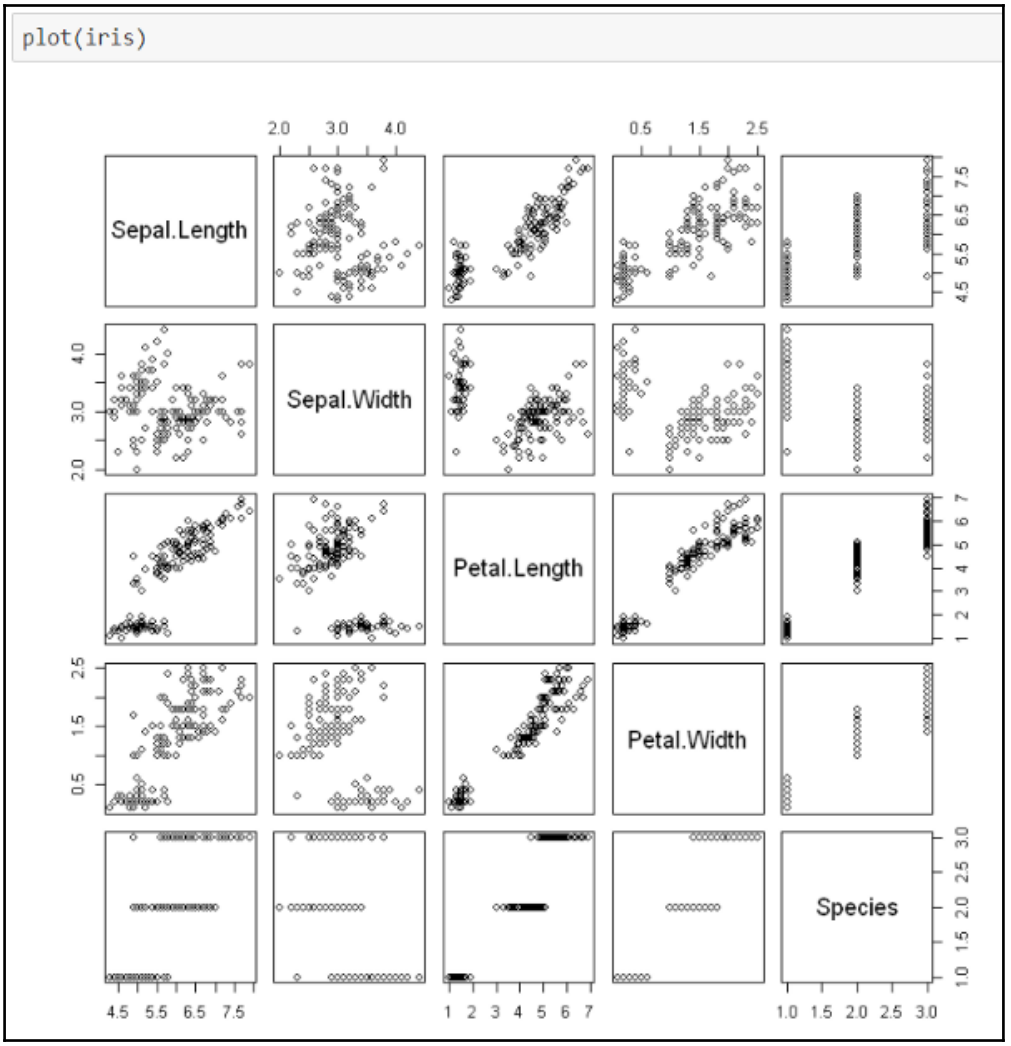

C:\Users\Dan\Downloads\B09656_07+r+iris+for+conversions.r
File Edit Search View Encoding Language Settings Tools
B09656_07+r+iris+for+conversions.r
1
2 data(iris)
3 head(iris)
4
5 summary(iris)
6
7 plot(iris)
8


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
B09656_07+r+iris+for+conversions.r
Source on Save
Run
Source
1 data(iris)
2 head(iris)
3
4 summary(iris)
5
6 plot(iris)
7
8
1:1
(Top Level)
R Script
Console ~/
> head(iris)
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1 5.1 3.5 1.4 0.2 setosa
2 4.9 3.0 1.4 0.2 setosa
3 4.7 3.2 1.3 0.2 setosa
4 4.6 3.1 1.5 0.2 setosa
5 5.0 3.6 1.4 0.2 setosa
6 5.4 3.9 1.7 0.4 setosa
> summary(iris)
Sepal.Length Sepal.Width Petal.Length
Min. :4.300 Min. :2.000 Min. :1.000
1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600
Median :5.800 Median :3.000 Median :4.350
Mean :5.843 Mean :3.057 Mean :3.758
3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100
Max. :7.900 Max. :4.400 Max. :6.900
Petal.Width Species
Min. :0.100 setosa :50
1st Qu.:0.300 versicolor:50
Median :1.300 virginica :50
Mean :1.199
3rd Qu.:1.800
Max. :2.500
> plot(iris)


R Graphics: Device 2 (ACTIVE)
File History Resize
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species


B09656_07 r iris for conv
file:///C:/Users/Dan/Downloads/B09656_07+r+iris+for+conversions.html
In [4]:
data(iris)
head(iris)
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
5.1 3.5 1.4 0.2 setosa
4.9 3.0 1.4 0.2 setosa
4.7 3.2 1.3 0.2 setosa
4.6 3.1 1.5 0.2 setosa
5.0 3.6 1.4 0.2 setosa
5.4 3.9 1.7 0.4 setosa
In [5]:
summary(iris)
Sepal.Length Sepal.Width Petal.Length Petal.Width
Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100
1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300
Median :5.800 Median :3.000 Median :4.350 Median :1.300
Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199
3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800
Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500
Species
setosa :50
versicolor:50
virginica :50

```
plot(iris)
```



```
11841 <div class="output_text output_subarea ">
11842 <pre>  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
11843  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
11844  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
11845  Median :5.800   Median :3.000   Median :4.350   Median :1.300
11846  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
11847  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
11848  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
11849        Species
11850  setosa     :50
11851  versicolor:50
11852  virginica :50
11853
11854
11855                  </pre>
11885
11886
11887 <div class="output_png output_subarea ">
11888 <img src="
11889 jIyampqnp6eysrK9vb3Hx8fQ0NDZ2dnh4eHp6enw8PD////QFLu4AAAACXBIWXMAABJ0AAAS
11890 dAHeZh94AAAgAElEQVR4nO1diZajOAx0n9PbPdPdP99uEsCnL8CBKl6uxkuS4tvQtmmtVmvV
11891 AGAY5mgCAHAFQEgAIAICQAEACEBgAAgJAAQAIQEAAAUBIACAAkAhQgIAAAAAAAAAAAAAAAA
11892 BAAhAYAAICQAEACEBAACgJAAQAAQEgAIAEICAAFASAAgAAgJAAQAIQGAACAkABAAhAQAAoCQ
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
```

```R
data(iris)
head(iris)
```
<table>
<thead><tr><th scope=col>Sepal.Length</th><th scope=col>Sepal.Width</th><th scope=col>Peta
<tbody>
    <tr><td>5.1   </td><td>3.5   </td><td>1.4   </td><td>0.2   </td><td>setosa</td></tr>
    <tr><td>4.9   </td><td>3.0   </td><td>1.4   </td><td>0.2   </td><td>setosa</td></tr>
    <tr><td>4.7   </td><td>3.2   </td><td>1.3   </td><td>0.2   </td><td>setosa</td></tr>
    <tr><td>4.6   </td><td>3.1   </td><td>1.5   </td><td>0.2   </td><td>setosa</td></tr>
    <tr><td>5.0   </td><td>3.6   </td><td>1.4   </td><td>0.2   </td><td>setosa</td></tr>
    <tr><td>5.4   </td><td>3.9   </td><td>1.7   </td><td>0.4   </td><td>setosa</td></tr>
</tbody>
</table>

```R
summary(iris)
```
```
      Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
     Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
     1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
     Median :5.800   Median :3.000   Median :4.350   Median :1.300
     Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
     3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
     Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
           Species
     setosa    :50
     versicolor:50
     virginica :50
```
```R
plot(iris)
```
```
![png](output_2_0.png)
```

```
data(iris)
head(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

```
summary(iris)
```

```
 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
      Species
setosa    :50
versicolor:50
```

```
.. code:: r

    data(iris)
    head(iris)

.. raw:: html

    <table>
    <thead><tr><th scope=col>Sepal.Length</th><th scope=col>Sepal.Width</th><th
scope=col>Petal.Length</th><th scope=col>Petal.Width</th><th scope=col>Species</th>
</tr></thead>
    <tbody>
        <tr><td>5.1    </td><td>3.5    </td><td>1.4    </td><td>0.2
</td><td>setosa</td></tr>
        <tr><td>4.9    </td><td>3.0    </td><td>1.4    </td><td>0.2
</td><td>setosa</td></tr>
        <tr><td>4.7    </td><td>3.2    </td><td>1.3    </td><td>0.2
</td><td>setosa</td></tr>
        <tr><td>4.6    </td><td>3.1    </td><td>1.5    </td><td>0.2
</td><td>setosa</td></tr>
        <tr><td>5.0    </td><td>3.6    </td><td>1.4    </td><td>0.2
</td><td>setosa</td></tr>
        <tr><td>5.4    </td><td>3.9    </td><td>1.7    </td><td>0.4
</td><td>setosa</td></tr>
    </tbody>
    </table>
.. code:: r

    summary(iris)
```

B09656_07 r iris for conversions

January 1, 2018

```
In [4]: data(iris)
        head(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

```
In [5]: summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

```
281
282
283     \begin{Verbatim}[commandchars=\\\{\}]
284 {\color{incolor}In [{\color{incolor}4}]:} data\PY{p}{()}iris\PY{p}{()}
285         \PY{k+kp}{head}\PY{p}{()}iris\PY{p}{()}
286 \end{Verbatim}
287
288
289     \begin{tabular}{r|lllll}
290  Sepal.Length & Sepal.Width & Petal.Length & Petal.Width & Species\\
291 \hline
292      5.1    & 3.5    & 1.4    & 0.2    & setosa\\
293      4.9    & 3.0    & 1.4    & 0.2    & setosa\\
294      4.7    & 3.2    & 1.3    & 0.2    & setosa\\
295      4.6    & 3.1    & 1.5    & 0.2    & setosa\\
296      5.0    & 3.6    & 1.4    & 0.2    & setosa\\
297      5.4    & 3.9    & 1.7    & 0.4    & setosa\\
298 \end{tabular}
299
300
301
302     \begin{Verbatim}[commandchars=\\\{\}]
303 {\color{incolor}In [{\color{incolor}5}]:} \PY{k+kp}{summary}\PY{p}{()}iris\PY{p}{()}
304 \end{Verbatim}
305
306
307
308     \begin{verbatim}
309   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
310  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
311  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
312  Median :5.800   Median :3.000   Median :4.350   Median :1.300
313  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
314  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
315  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
316        Species
317  setosa    :50
318  versicolor:50
319  virginica :50
320
321
322
323     \end{verbatim}
324
325
326     \begin{Verbatim}[commandchars=\\\{\}]
327 {\color{incolor}In [{\color{incolor}6}]:} plot\PY{p}{()}iris\PY{p}{()}
328 \end{Verbatim}
329
```

500 : Internal Server Error

The error was:

```
nbconvert failed: xelatex not found on PATH, if you have not installed xelatex you may need to do so. Find fu
rther instructions at https://nbconvert.readthedocs.io/en/latest/install.html#installing-tex.
```

B09656_07 r iris for conversions

January 1, 2018

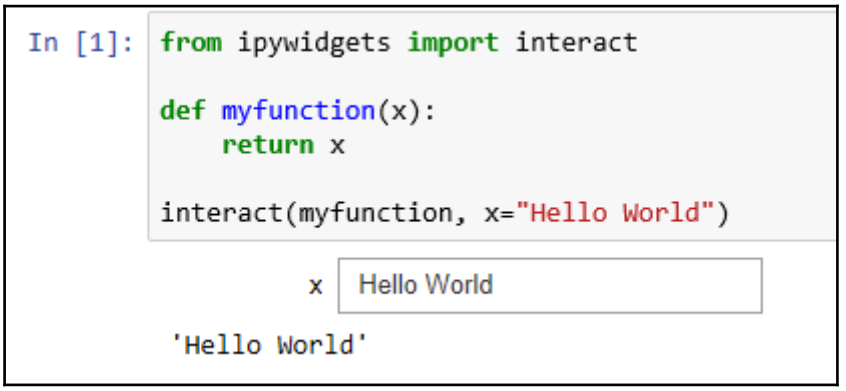
```
In [4]: data(iris)
        head(iris)
```

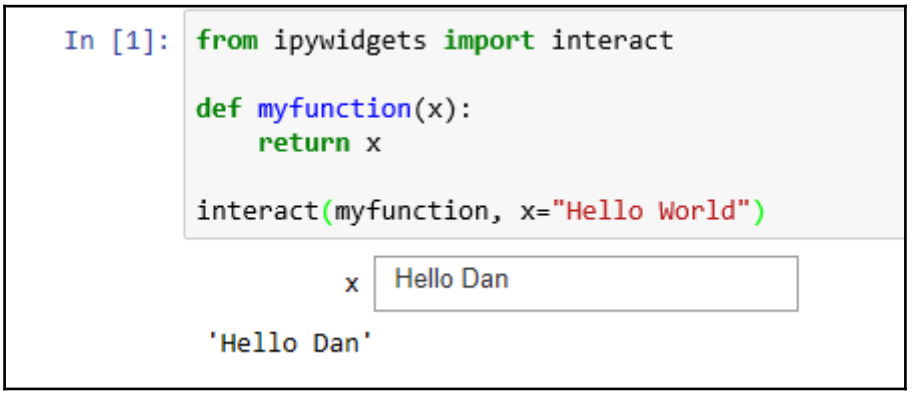
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

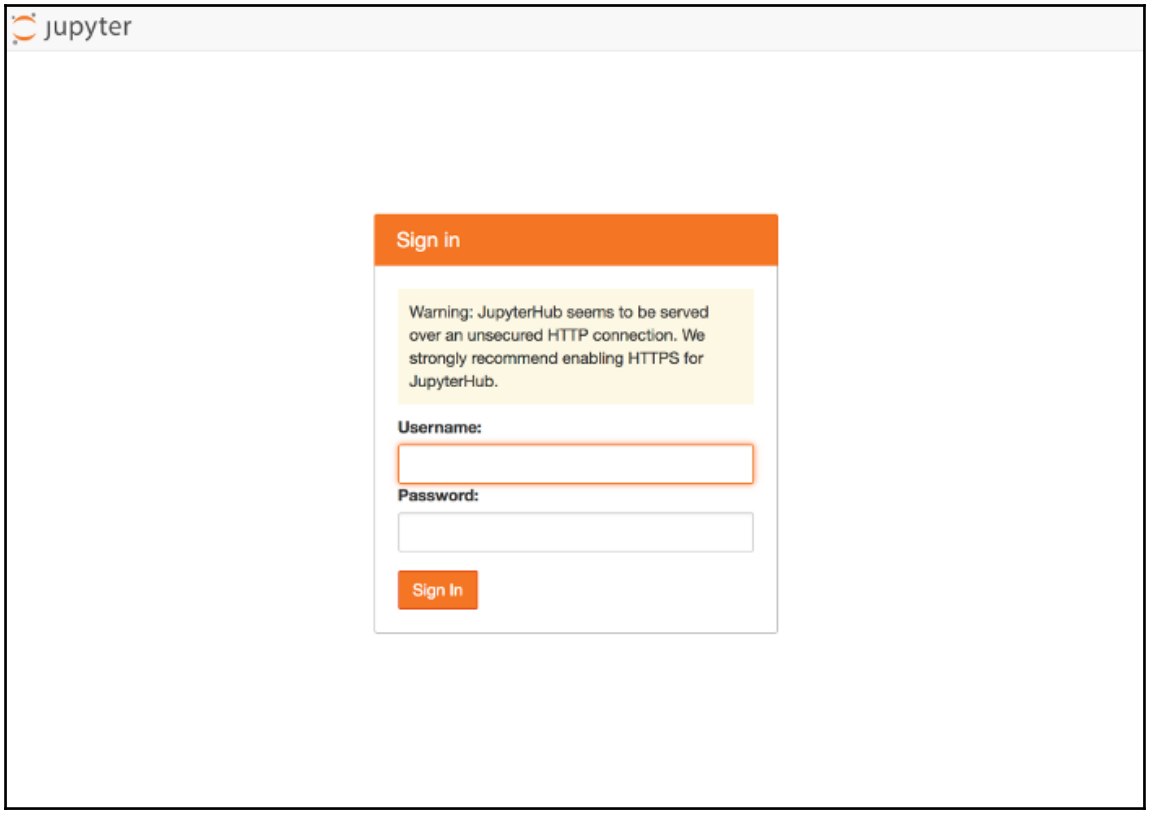
```
In [5]: summary(iris)
```

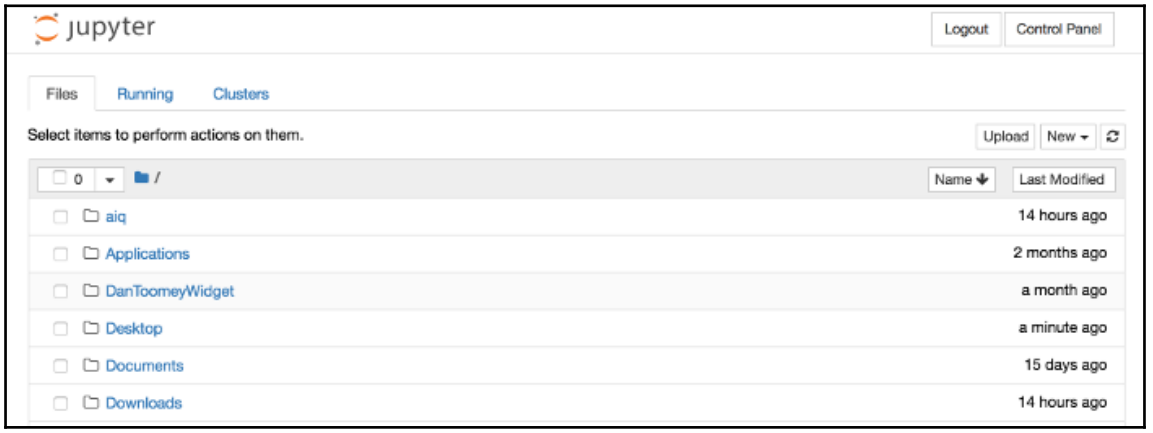
```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa     :50
 versicolor:50
 virginica :50
```

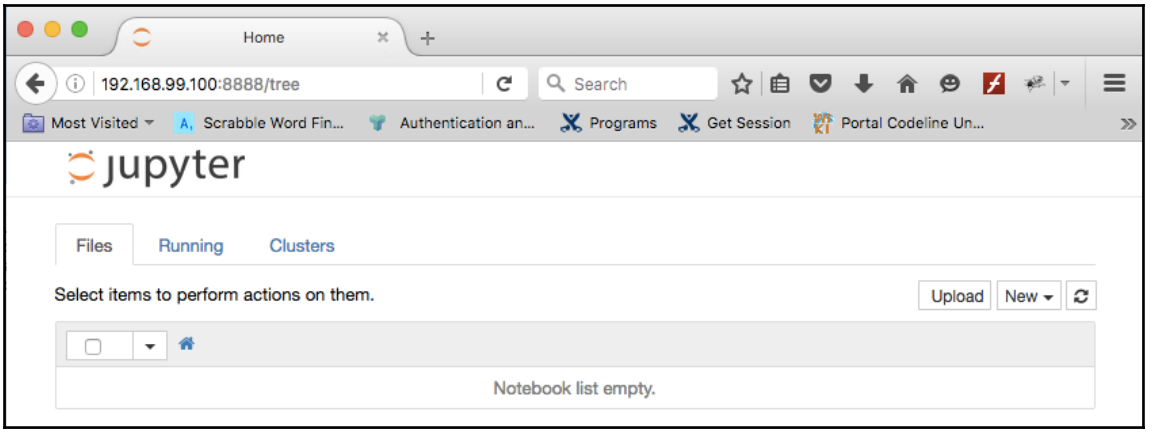
Chapter 8: Multiuser Jupyter






jupyter
Sign in
Warning: JupyterHub seems to be served over an unsecured HTTP connection. We strongly recommend enabling HTTPS for JupyterHub.
Username:
Password:
Sign In


jupyter
Logout
Control Panel
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
0
/
Name
Last Modified
aiq
14 hours ago
Applications
2 months ago
DanToomeyWidget
a month ago
Desktop
a minute ago
Documents
15 days ago
Downloads
14 hours ago


Home
192.168.99.100:8888/tree
Search
Most Visited
Scrabble Word Fin...
Authentication an...
Programs
Get Session
Portal Codeline Un...
jupyter
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
Notebook list empty.

Chapter 9: Interacting with Big Data

```
In [9]: import pyspark
        if not 'sc' in globals():
            sc = pyspark.SparkContext()

        text_file = sc.textFile('B09656_09_word_count.ipynb')
        counts = text_file.flatMap(lambda line: line.split(" ")) \
                    .map(lambda word: (word, 1)) \
                    .reduceByKey(lambda a, b: a + b)

        for x in counts.collect():
            print (x)
```

```
('', 291)
('"cells":', 1)
('[', 4)
('"code",', 1)
('8,', 1)
('"outputs":', 1)
('"NameError",', 1)
('"evalue":', 1)
('"name', 1)
```

In [1]:

```
import pyspark
if not 'sc' in globals():
    sc = pyspark.SparkContext()

text_file = sc.textFile('B09656_09_word_count.ipynb')
counts = text_file.flatMap(lambda line: line.split(" ")) \
             .map(lambda word: (word, 1)) \
             .reduceByKey(lambda a, b: a + b) \
             .sortByKey()

for x in counts.collect():
    print (x)
```

```
('', 929)
('"', 5)
('"(\'\',', 1)
('"(\'(most\',', 1)
('"(\'(word,\',', 1)
('"(\'(x)\\"\',', 1)
('"(\'+\',', 1)
('"(\'.map(lambda\',', 1)
('"(\'.reduceByKey(lambda\',', 1)
('"(\'1))\',', 1)
```

In [3]:

```
import pyspark
if not 'sc' in globals():
    sc = pyspark.SparkContext()

textFile = sc.textFile('access_log')
print(textFile.count(),"access records")

gets = textFile.filter(lambda line: "GET" in line)
print(gets.count(),"GETs")

posts = textFile.filter(lambda line: "POST" in line)
print(posts.count(),"POSTs")

other = textFile.subtract(gets).subtract(posts)
print(other.count(),"Other")

for x in other.collect():
    print(x)
```

```
1546 access records
1525 GETs
14 POSTs
7 Other
h194n2fls308o1033.telia.com - - [09/Mar/2004:13:49:05 -0800] "-" 408 -
64.246.94.141 - - [10/Mar/2004:16:31:19 -0800] "HEAD /twiki/bin/view/Main/SpamAssassinDeletin
g HTTP/1.1" 200 0
206-15-133-154.dialup.ziplink.net - - [11/Mar/2004:16:33:23 -0800] "HEAD /twiki/bin/view/Mai
n/SpamAssassinDeleting HTTP/1.1" 200 0
```

In [1]:

```
import pyspark
if not 'sc' in globals():
    sc = pyspark.SparkContext()

#check if a number is prime
def isprime(n):
    # must be positive
    n = abs(int(n))

    # 2 or more
    if n < 2:
        return False

    # 2 is the only even prime number
    if n == 2:
        return True
    if not n & 1:
        return False

    # range starts with 3 and only needs to go up the square root of n
    # for all odd numbers
    for x in range(3, int(n**0.5)+1, 2):
        if n % x == 0:
            return False
    return True

nums = sc.parallelize(range(1000000))

# Compute the number of primes in the RDD
print(nums.filter(isprime).count())
```

```
78498
```

In [1]:
```
import pyspark
if not 'sc' in globals():
    sc = pyspark.SparkContext()

sentences = sc.textFile('B09656_09_article.txt') \
              .glom() \
              .map(lambda x: " ".join(x)) \
              .flatMap(lambda x: x.split("."))
print(sentences.count(),"sentences")

bigrams = sentences.map(lambda x:x.split()) \
              .flatMap(lambda x: [((x[i],x[i+1]),1) for i in range(0,len(x)-1)])
print(bigrams.count(),"bigrams")

frequent_bigrams = bigrams.reduceByKey(lambda x,y:x+y) \
              .map(lambda x:(x[1],x[0])) \
              .sortByKey(False)
frequent_bigrams.take(10)
```

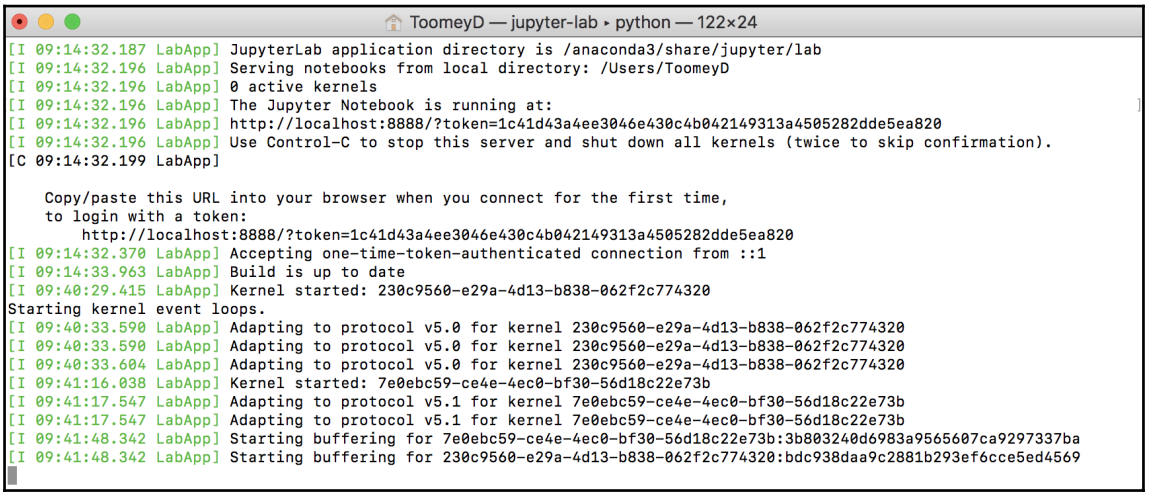
```
140 sentences
2448 bigrams
```

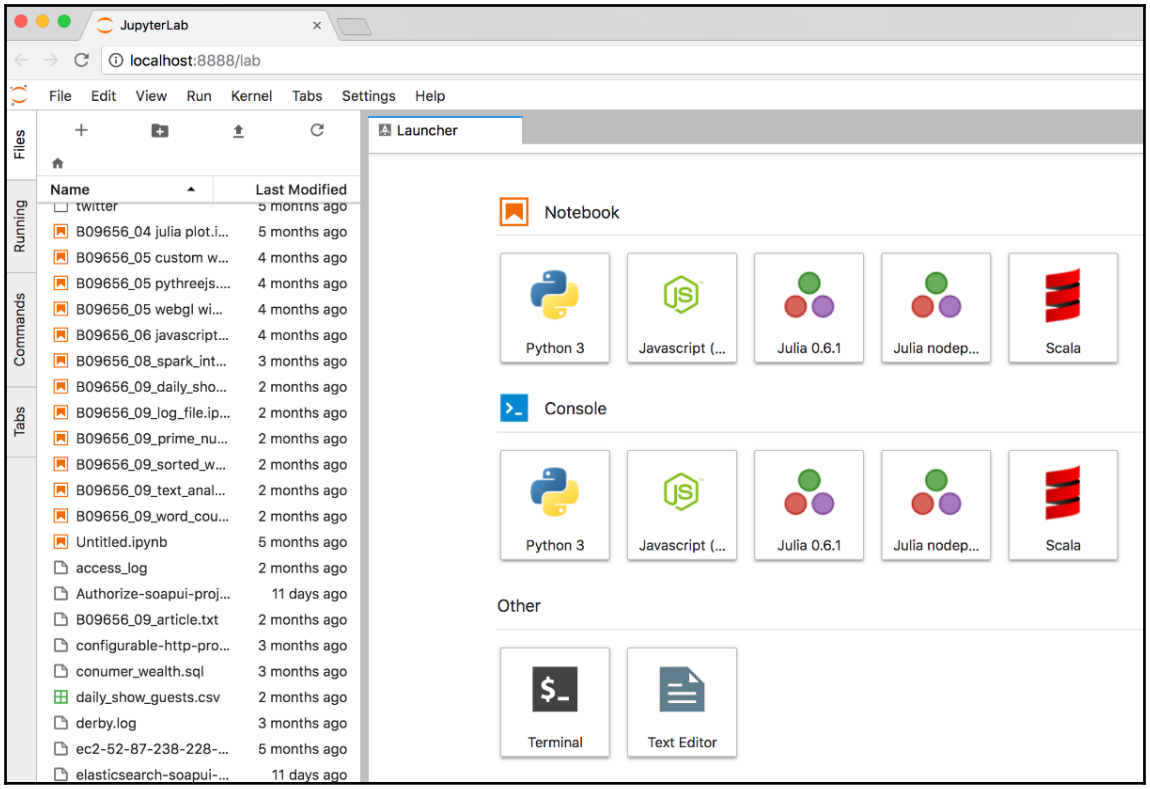
Out[1]:
```
[(11, ('of', 'the')),
 (8, ('in', 'the')),
 (6, ('to', 'the')),
 (6, ('for', 'the')),
 (6, ('of', 'college')),
 (5, ('more', 'than')),
 (5, ('I\ufffdm', 'cynical')),
 (5, ('cynical', 'about')),
 (5, ('The', 'vast')),
 (5, ('vocational', 'education'))]
```

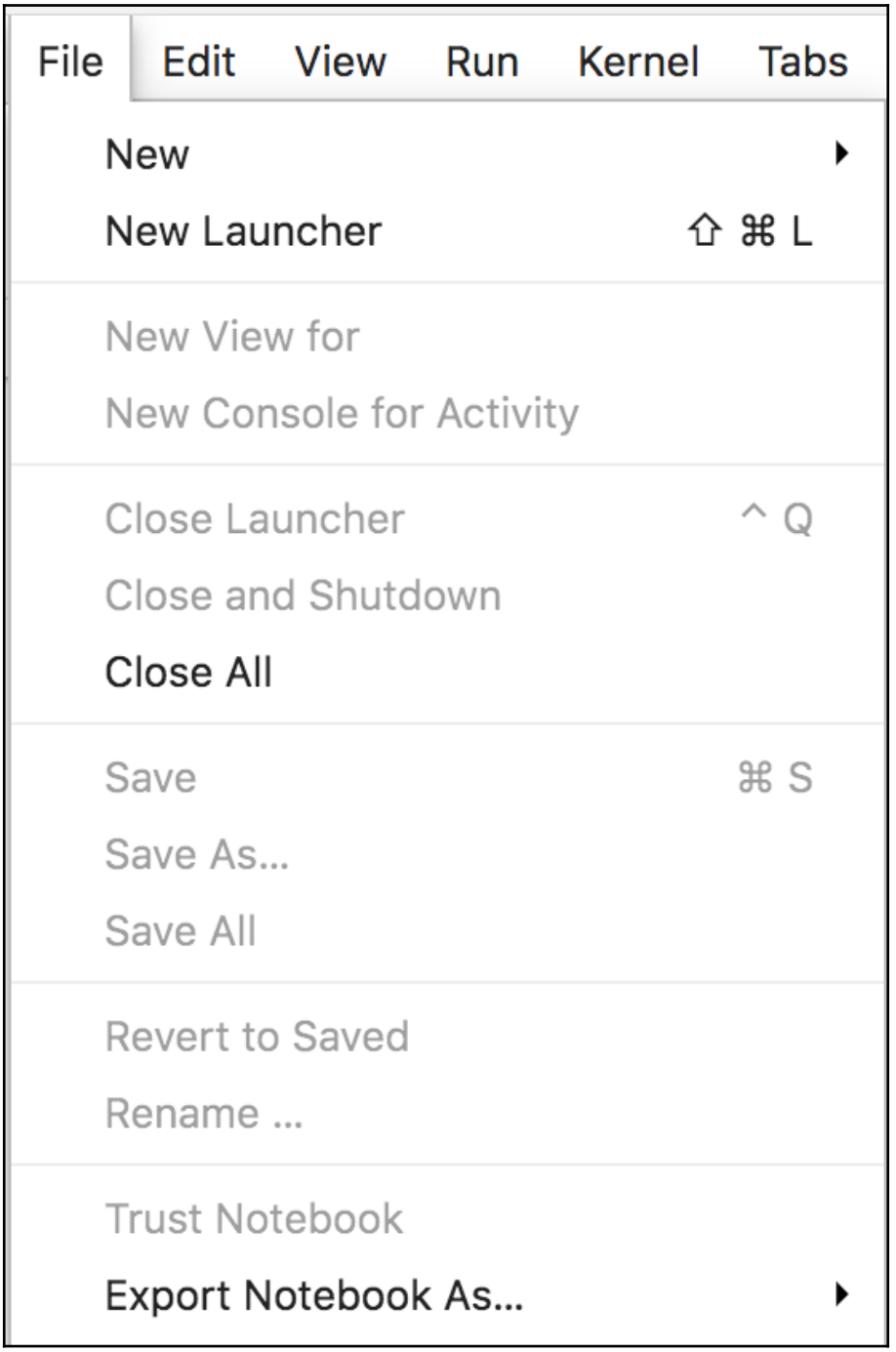
```
syears = sorted(years.items(), key=operator.itemgetter(1), reverse=True)
soccupations = sorted(occupations.items(), key=operator.itemgetter(1), reverse=True)
sguests = sorted(guests.items(), key=operator.itemgetter(1), reverse=True)

print(syears[:5])
print(soccupations[:5])
print(sguests[:5])
```

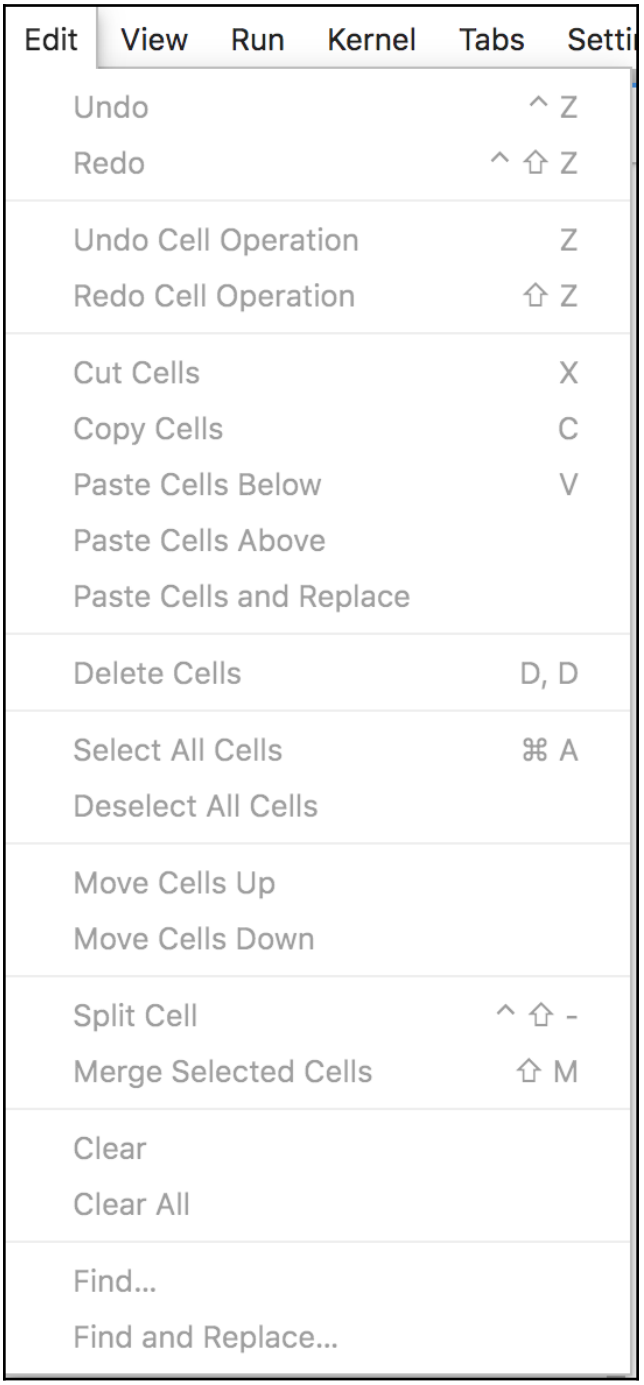
```
got error
[('2000', 169), ('1999', 166), ('2001', 16)]
[('actor', 122), ('actress', 73), ('comedian', 19), ('film actress', 7), ('NA', 7)]
[('Bob Dole', 5), ('"Pamela Anderson', 3), ('Adam Sandler', 3), ('Richard Belzer', 3), ('Trac
ey Ullman', 2)]
```

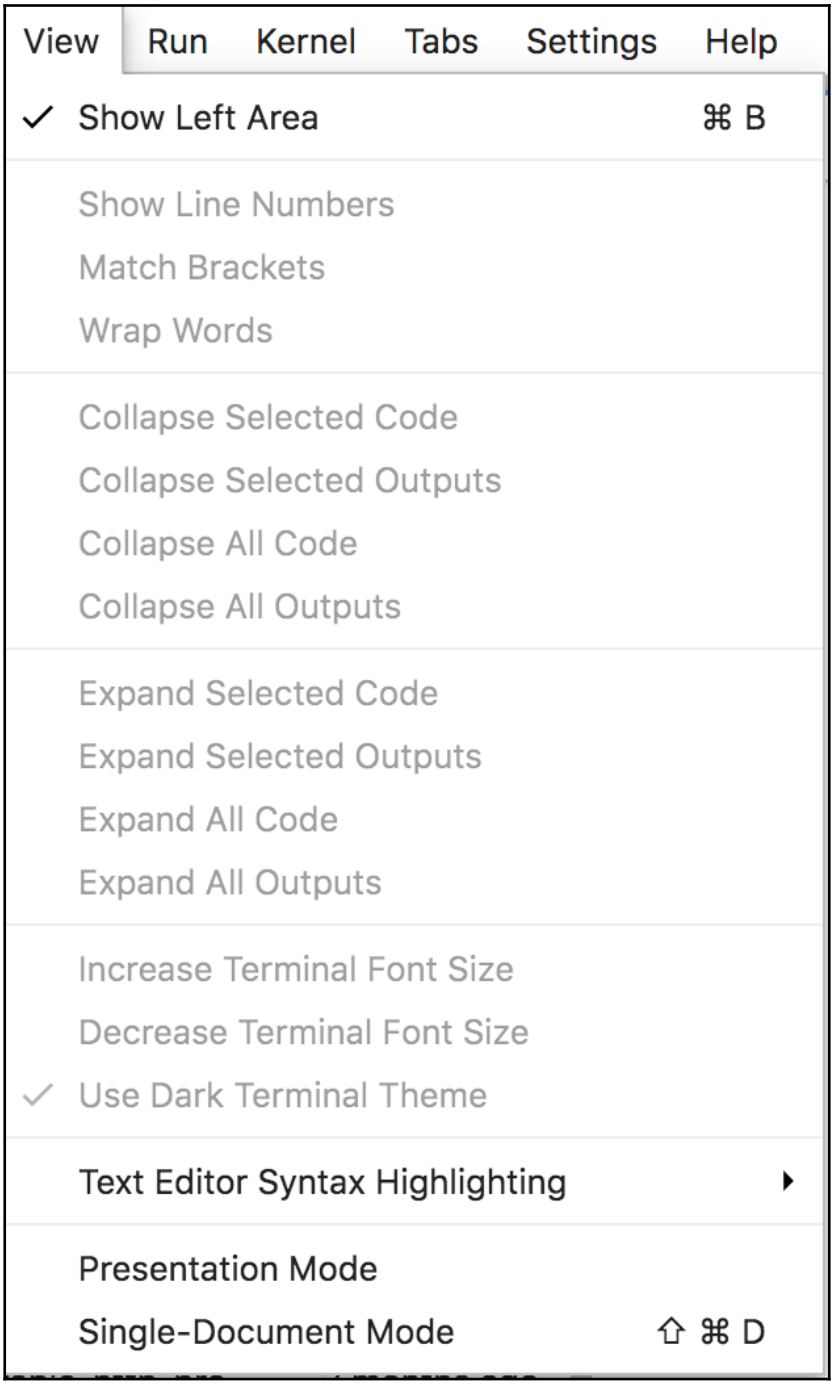
Chapter 11: Jupyter Labs

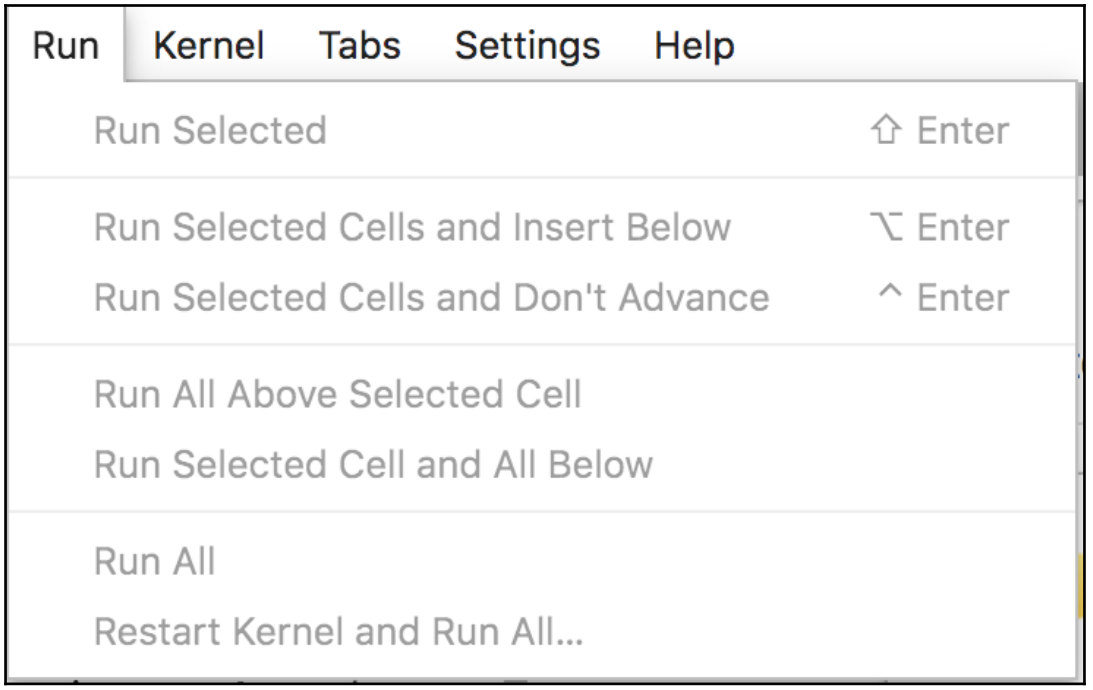


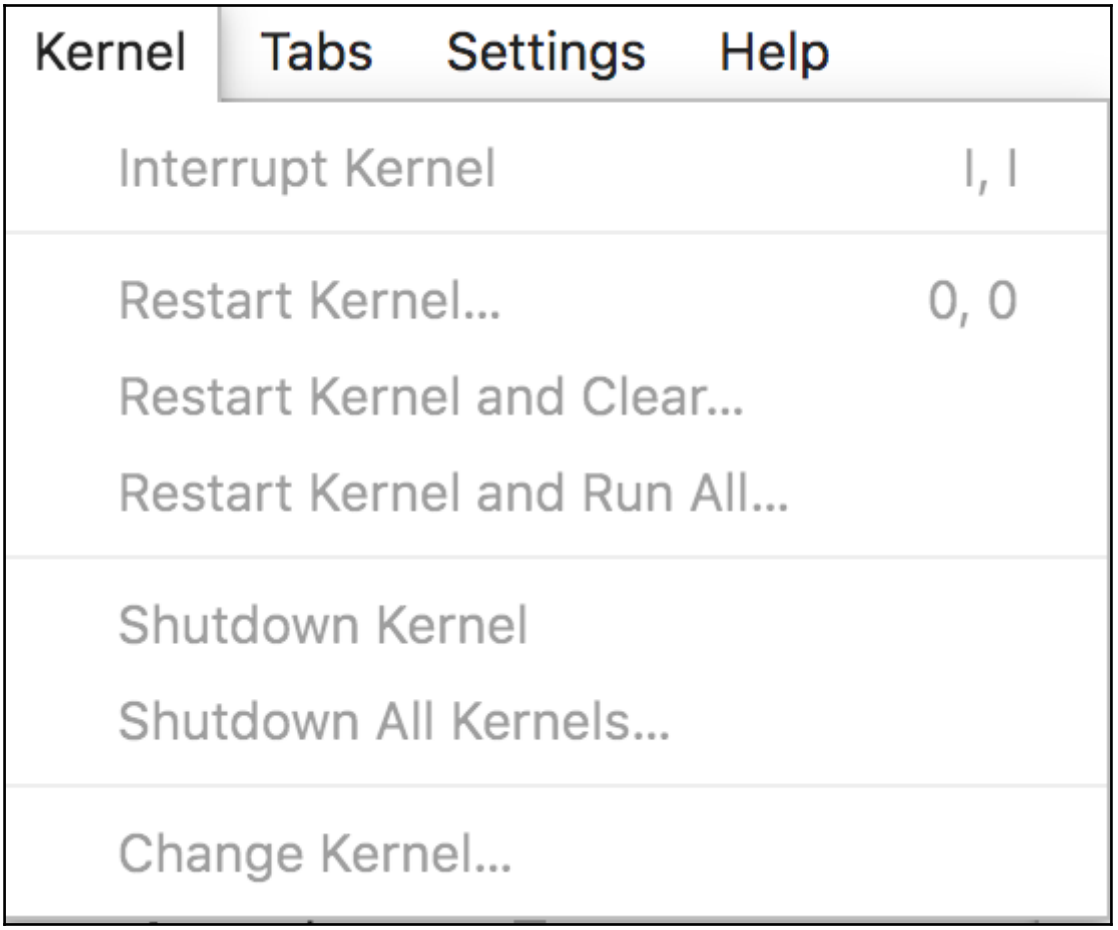

JupyterLab
localhost:8888/lab
File Edit View Run Kernel Tabs Settings Help
Files
Running
Commands
Tabs
Name
Last Modified
B09656_04 julia plot.i... 5 months ago
B09656_05 custom w... 4 months ago
B09656_05 pythreejs.... 4 months ago
B09656_05 webgl wi... 4 months ago
B09656_06 javascript... 4 months ago
B09656_08_spark_int... 3 months ago
B09656_09_daily_sho... 2 months ago
B09656_09_log_file.ip... 2 months ago
B09656_09_prime_nu... 2 months ago
B09656_09_sorted_w... 2 months ago
B09656_09_text_anal... 2 months ago
B09656_09_word_cou... 2 months ago
Untitled.ipynb 5 months ago
access_log 2 months ago
Authorize-soapui-proj... 11 days ago
B09656_09_article.txt 2 months ago
configurable-http-pro... 3 months ago
conumer_wealth.sql 3 months ago
daily_show_guests.csv 2 months ago
derby.log 3 months ago
ec2-52-87-238-228-... 5 months ago
elasticsearch-soapui-... 11 days ago
Launcher
Notebook
Python 3
Javascript (...
Julia 0.6.1
Julia nodep...
Scala
Console
Python 3
Javascript (...
Julia 0.6.1
Julia nodep...
Scala
Other
Terminal
Text Editor

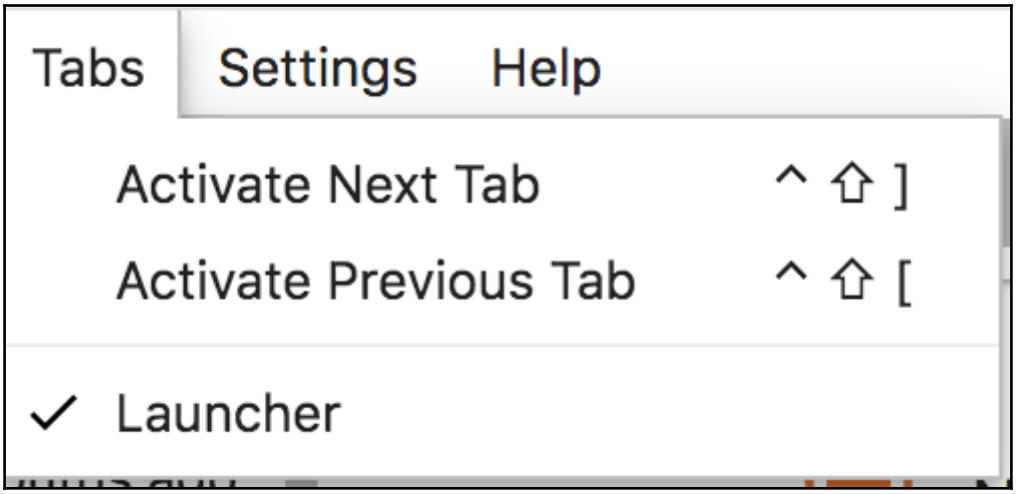

File
Edit
View
Run
Kernel
Tabs
New
New Launcher
⇧ ⌘ L
New View for
New Console for Activity
Close Launcher
^ Q
Close and Shutdown
Close All
Save
⌘ S
Save As...
Save All
Revert to Saved
Rename ...
Trust Notebook
Export Notebook As...

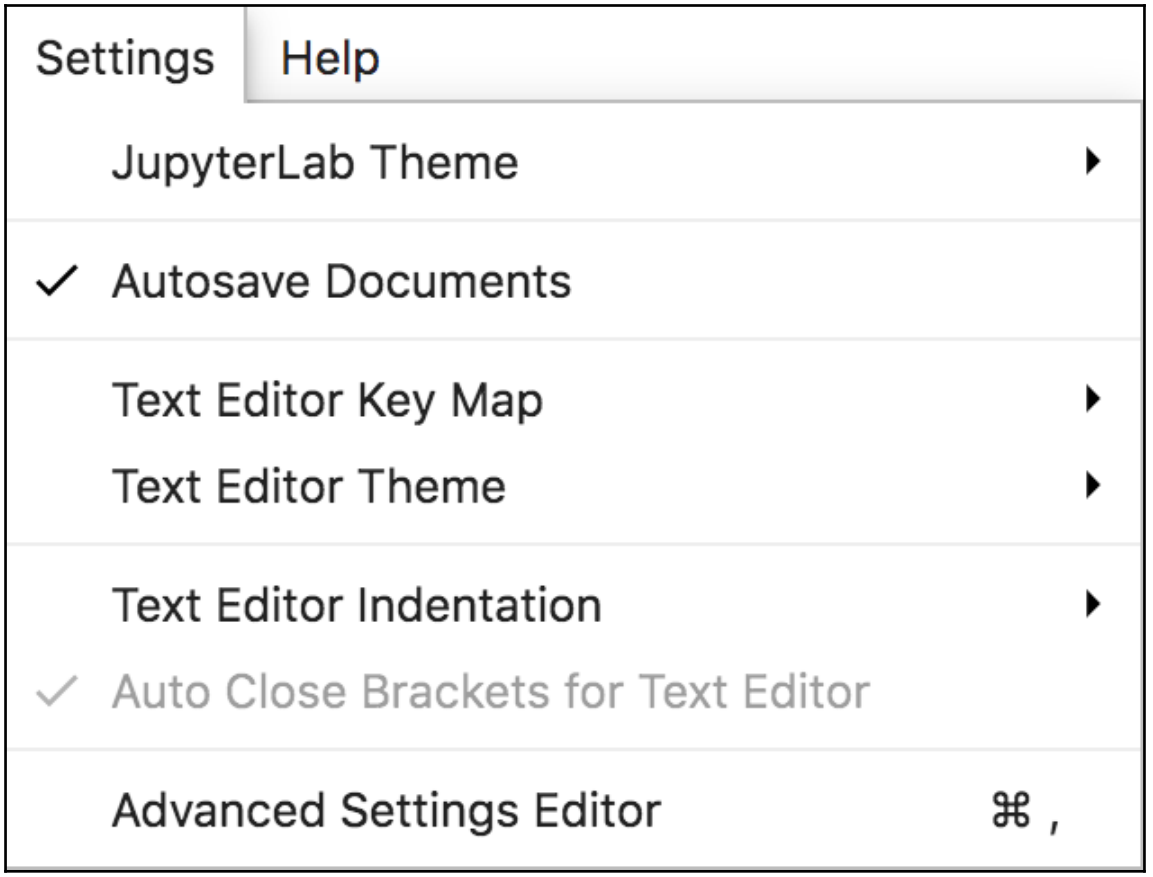

Edit
View
Run
Kernel
Tabs
Setti
Undo ^ Z
Redo ^ ⇧ Z
Undo Cell Operation Z
Redo Cell Operation ⇧ Z
Cut Cells X
Copy Cells C
Paste Cells Below V
Paste Cells Above
Paste Cells and Replace
Delete Cells D, D
Select All Cells ⌘ A
Deselect All Cells
Move Cells Up
Move Cells Down
Split Cell ^ ⇧ -
Merge Selected Cells ⇧ M
Clear
Clear All
Find...
Find and Replace...

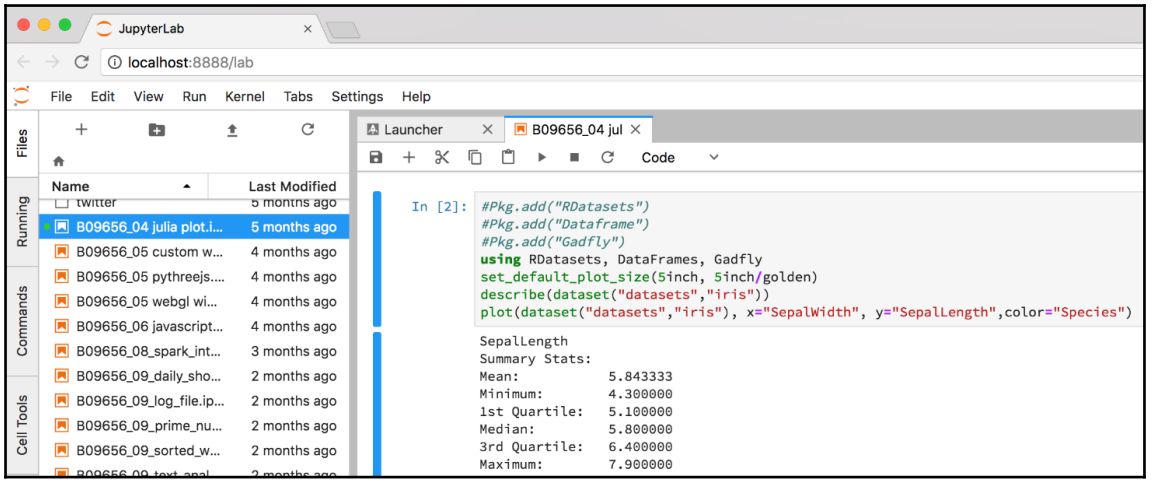

View
Run
Kernel
Tabs
Settings
Help
Show Left Area
⌘ B
Show Line Numbers
Match Brackets
Wrap Words
Collapse Selected Code
Collapse Selected Outputs
Collapse All Code
Collapse All Outputs
Expand Selected Code
Expand Selected Outputs
Expand All Code
Expand All Outputs
Increase Terminal Font Size
Decrease Terminal Font Size
Use Dark Terminal Theme
Text Editor Syntax Highlighting
Presentation Mode
Single-Document Mode
⇧ ⌘ D

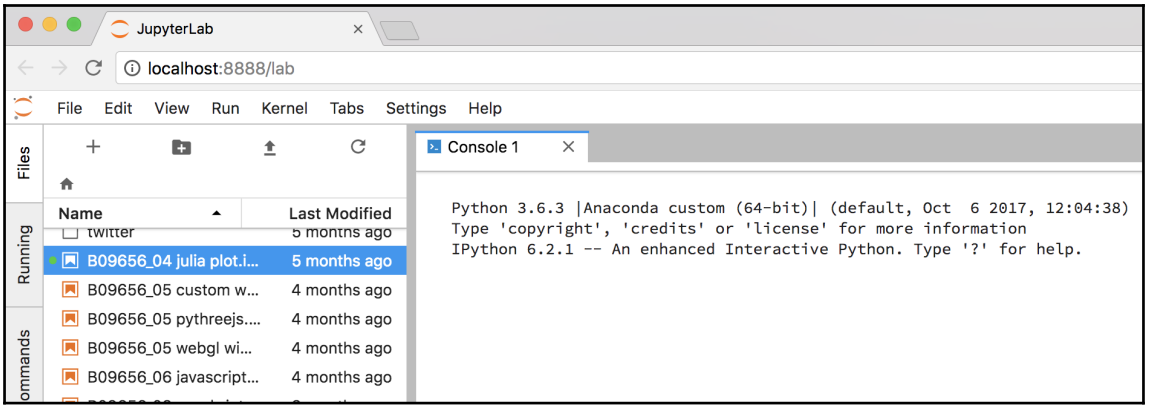

Run
Kernel
Tabs
Settings
Help
Run Selected
⇧ Enter
Run Selected Cells and Insert Below
⌥ Enter
Run Selected Cells and Don't Advance
^ Enter
Run All Above Selected Cell
Run Selected Cell and All Below
Run All
Restart Kernel and Run All...


Kernel
Tabs
Settings
Help
Interrupt Kernel
I, I
Restart Kernel...
0, 0
Restart Kernel and Clear...
Restart Kernel and Run All...
Shutdown Kernel
Shutdown All Kernels...
Change Kernel...


Tabs
Settings
Help
Activate Next Tab
^ ⇧]
Activate Previous Tab
^ ⇧ [
Launcher


Settings
Help
JupyterLab Theme
Autosave Documents
Text Editor Key Map
Text Editor Theme
Text Editor Indentation
Auto Close Brackets for Text Editor
Advanced Settings Editor
⌘ ,


JupyterLab
localhost:8888/lab
File Edit View Run Kernel Tabs Settings Help
Files
Running
Commands
Cell Tools
Name
Last Modified
B09656_04 julia plot.i... 5 months ago
B09656_05 custom w... 4 months ago
B09656_05 pythreejs.... 4 months ago
B09656_05 webgl wi... 4 months ago
B09656_06 javascript... 4 months ago
B09656_08_spark_int... 3 months ago
B09656_09_daily_sho... 2 months ago
B09656_09_log_file.ip... 2 months ago
B09656_09_prime_nu... 2 months ago
B09656_09_sorted_w... 2 months ago
Launcher
B09656_04 jul
Code
In [2]:
#Pkg.add("RDatasets")
#Pkg.add("Dataframe")
#Pkg.add("Gadfly")
using RDatasets, DataFrames, Gadfly
set_default_plot_size(5inch, 5inch/golden)
describe(dataset("datasets","iris"))
plot(dataset("datasets","iris"), x="SepalWidth", y="SepalLength",color="Species")
SepalLength
Summary Stats:
Mean: 5.843333
Minimum: 4.300000
1st Quartile: 5.100000
Median: 5.800000
3rd Quartile: 6.400000
Maximum: 7.900000


JupyterLab
localhost:8888/lab
File Edit View Run Kernel Tabs Settings Help
Files
Running
Commands
Name
Last Modified
twitter 5 months ago
B09656_04 julia plot.i... 5 months ago
B09656_05 custom w... 4 months ago
B09656_05 pythreejs.... 4 months ago
B09656_05 webgl wi... 4 months ago
B09656_06 javascript... 4 months ago
Console 1
Python 3.6.3 |Anaconda custom (64-bit)| (default, Oct 6 2017, 12:04:38)
Type 'copyright', 'credits' or 'license' for more information
IPython 6.2.1 -- An enhanced Interactive Python. Type '?' for help.