Introduction to R and Data Science Tools in the Microsoft Stack

Jamey Johnston

Agenda

- **Intro to R**
  - *R and RStudio*
  - *Basics*
  - *Objects in R*
  - *Packages*
  - *Control Flows*
  - *RStudio Overview*
- **MS and R**
  - *Databricks*
  - *Azure ML*
  - *MS Machine Learning Services*
  - *SQL 2016+*
  - *Power BI*
- **Resources**

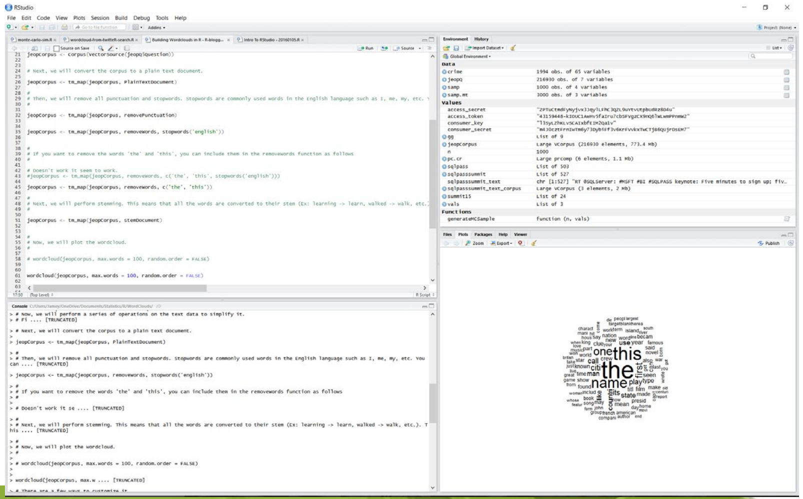
Source: https://www.r-project.org/logo/

Jamey Johnston

- Sr. Data Scientist/Engineer for O&G Company
- 25+ years Data Experience
- TAMU MS in Analytics
  - http://analytics.stat.tamu.edu
- Semi-Pro Photographer
  - http://jamey.photography
- @STATCowboy
- GitHub (code) - https://github.com/STATCowboy/CodeLIkePirate
- http://STATCowboy.com

R and RStudio

- R Project for Statistical Computing
  - https://www.r-project.org/
- RStudio
  - https://www.rstudio.com/



RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
monte-carlo-sim.R
wordcloud-from-twitteR-search.R
Building Wordclouds in R - R-blogg...
Intro To RStudio - 20160105.R
Source on Save
Run
Source
Project: (None)
Environment
History
Import Dataset
List
Global Environment
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
17:50
(Top Level)
R Script
Console C:/Users/Jamey/OneDrive/Documents/Statistics/R/WordClouds/

```
jeopCorpus <- Corpus(VectorSource(jeopQ$Question))


# Next, we will convert the corpus to a plain text document.

jeopCorpus <- tm_map(jeopCorpus, PlainTextDocument)

#
# Then, we will remove all punctuation and stopwords. Stopwords are commonly used words in the English language such as I, me, my, etc. Y
#

jeopCorpus <- tm_map(jeopCorpus, removePunctuation)


jeopCorpus <- tm_map(jeopCorpus, removeWords, stopwords('english'))


#
# If you want to remove the words 'the' and 'this', you can include them in the removeWords function as follows
#

# Doesn't work it seem to work.
#jeopCorpus <- tm_map(jeopCorpus, removeWords, c('the', 'this', stopwords('english')))

jeopCorpus <- tm_map(jeopCorpus, removeWords, c('the', 'this'))

#
# Next, we will perform stemming. This means that all the words are converted to their stem (Ex: learning -> learn, walked -> walk, etc.)
#

jeopCorpus <- tm_map(jeopCorpus, stemDocument)


#
# Now, we will plot the wordcloud.
#

# wordcloud(jeopCorpus, max.words = 100, random.order = FALSE)


wordcloud(jeopCorpus, max.words = 100, random.order = FALSE)
```

```
> # Now, we will perform a series of operations on the text data to simplify it.
> # Fi .... [TRUNCATED]

> # Next, we will convert the corpus to a plain text document.
>
> jeopCorpus <- tm_map(jeopCorpus, PlainTextDocument)

> #
> # Then, we will remove all punctuation and stopwords. Stopwords are commonly used words in the English language such as I, me, my, etc. You
can .... [TRUNCATED]

> jeopCorpus <- tm_map(jeopCorpus, removeWords, stopwords('english'))

> #
> # If you want to remove the words 'the' and 'this', you can include them in the removeWords function as follows
> #
>
> # Doesn't work it se .... [TRUNCATED]

> #
> # Next, we will perform stemming. This means that all the words are converted to their stem (Ex: learning -> learn, walked -> walk, etc.). T
his .... [TRUNCATED]

> #
> # Now, we will plot the wordcloud.
> #
>
> # wordcloud(jeopCorpus, max.words = 100, random.order = FALSE)
>
>
> wordcloud(jeopCorpus, max.w .... [TRUNCATED]

> # There are a few ways to customize it
```

Data	
crime	1994 obs. of 65 variables
jeopQ	216930 obs. of 7 variables
samp	1000 obs. of 4 variables
samp.mt	3000 obs. of 3 variables
Values	
access_secret	"ZPTuCtmdFyNyjvx3JqylLFhC3QZL9uYtvUtpbudRz8D4u"
access_token	"43159448-kIOUC1AwHv5faIru7cbSFVgzCX9nQ6lWLwmPPnmW2"
consumer_key	"l3SyLZhKLvSCAIxbfEIM2Qa1v"
consumer_secret	"m4J0cztFrnIwTm6y73Dyb1Ff3v6KrFVvkxTWCTjB6QUjrDsEM7"
gg	List of 9
jeopCorpus	Large VCorpus (216930 elements, 773.4 Mb)
n	1000
pc.cr	Large prcomp (6 elements, 1.1 Mb)
sqlpass	List of 503
sqlpasssummit	List of 527
sqlpasssummit_text	chr [1:527] "RT @SQLServer: #MSFT #BI #SQLPASS keynote: Five minutes to sign up; fiv…
sqlpasssummit_text_corpus	Large VCorpus (3 elements, 2 Mb)
summit15	List of 24
vals	List of 3
Functions	
generateMCSample	function (n, vals)

Basics

- # - comment

```
> # Basics
```

- Variable Creation

```
> m <- 3 * 5
> m
[1] 15
```

- Help

```
> help(“lm”)  # lm is function for Fitting Linear Models
> ?lm
> lm(y ~ x)
```

Objects in R

- Variables, Values, Commands, Functions …
- Everything in R is an Object
- Typical Data in R is stored in:
  - Vectors (one row, same data type)
  - Matrices (multiple rows, same data type)
  - Data Frames (multiple rows, multiple data types)
    - It’s like a Table!
  - List (collection of objects)

Vector

- Building Blocks for data objects in R
- *c* (combine) function to create a Vector
  - `v <- c(2, 3, 1.5, 3.1, 49)`
- *seq* function generates numeric sequences
  - `s <- seq(from = 0, to = 100, by = .1)`
- *rep* function replicates values
  - `r <- rep(c(1,4), times = 4)`
- *:* creates a number seq incremented by 1 or -1
  - `colon <- 1:10`
- length(var) returns length of vector
  - `length(colon)`

Matrix

- *matrix* function used to build matrix
- rbind (row bind) and cbind (column bind)
  - Combine matrices by row or column
- http://www.ats.ucla.edu/stat/r/library/matrix_alg.htm
- Demos

Data Frame

- It is like a table!
- *rownames* – extract row labels
- *colnames* – extract column labels
- *read.table, read.csv, readxl, RODBC*
  - Different ways to create data frames
- Demos

List

- Combine multiple objects types into one object
    - vectors, matrices, data frames, list, functions
- Typically used by functions to output the model output
    - e.g. the output from the lm function
- Demo

Missing Data

- NA is used to represent Missing Data
- The *is.na* and *which* functions are used to manage NA

```
> x <- c(1.3,2.3,3.4,NA)
> print(x)
[1] 1.3 2.3 3.4  NA
>
> # Returns integer location of values (not the values)
> n <- which (is.na(x))
> v <- which (!is.na(x))
> print(n)
[1] 4
> print(v)
[1] 1 2 3
>
> # y will be set to the values not = NA
> y <- x[!is.na(x)]
> print(y)
[1] 1.3 2.3 3.4
```

Packages

- Add-ons for R
- *library()*
  - List packages already installed
- *install.package("dplyr2", "ggplot2")*
  - Install new packages
- *library(dplyr2)*
  - Load package to be used in R

Conditional Operators

- Comparisons return logical vector

```
> 1:10 == 2
 [1] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> 1:10 != 2
 [1]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> 1:10 > 2
 [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> 1:10 >= 2
 [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> 1:10 < 2
 [1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> 1:10 <= 2
 [1]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

> x <- 2
> x > 1
[1] TRUE
```

Logical Operations

```
> x <- 1:4
> x
[1] 1 2 3 4
>
> (x > 2) | (x <= 3)
[1] TRUE TRUE TRUE TRUE
>
> (x > 2) & (x <= 3)
[1] FALSE FALSE  TRUE FALSE
>
> xor((x > 2), (x < 4))
[1]  TRUE  TRUE FALSE  TRUE
>
> 0:5 %in% x
[1] FALSE  TRUE  TRUE  TRUE  TRUE FALSE
```

Control Flows

- IF … ELSE

```
x <- 4
if (x < 3) print("true") else print("false")
ifelse ((x < 3), print("true"), print("false"))
```

- FOR Loops

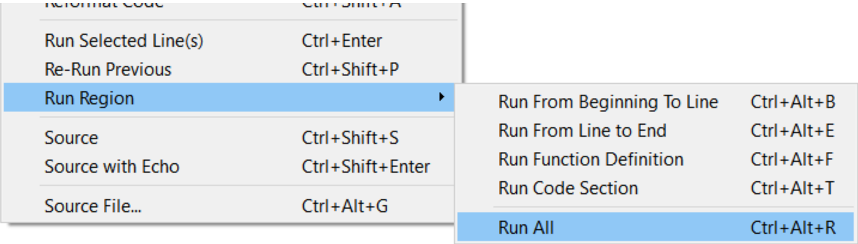
```
for(i in 1:10)
  print(1:i)

for (i in 1:nrow(df))
  print(df[i,])
```

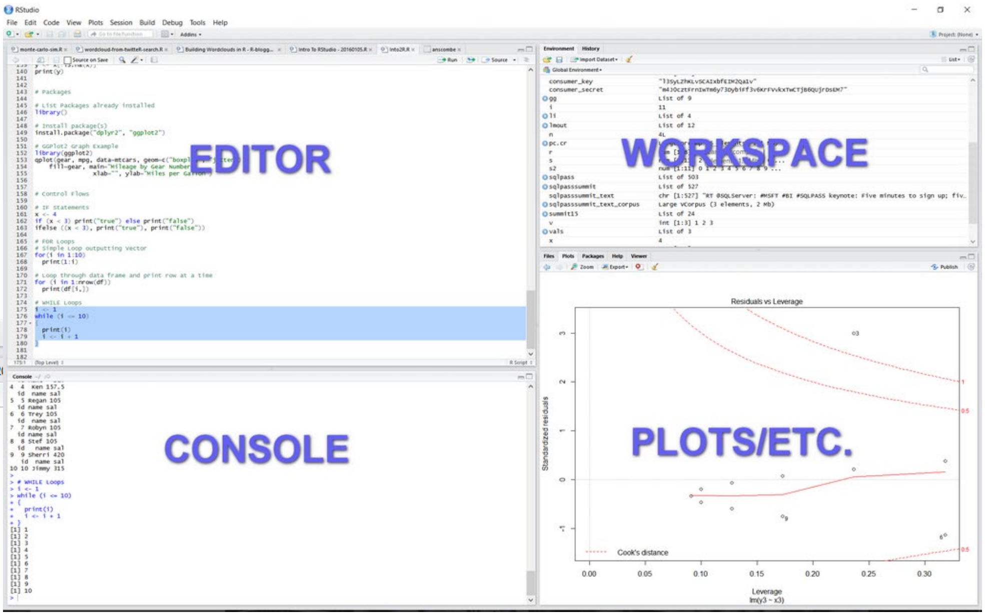
- WHILE Loops

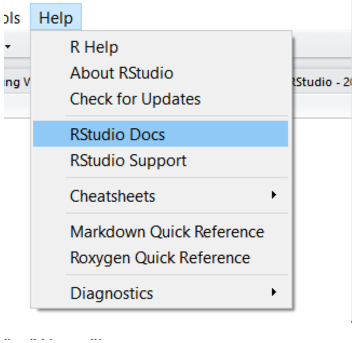
```
i <- 1
while (i <= 10)
{
  print(i)
  i <- i + 1
}
```

RStudio

- Run Options
  - CTL+Enter
  - Ctl+Alt+R
- Built-In Docs
- Version Control
- Projects

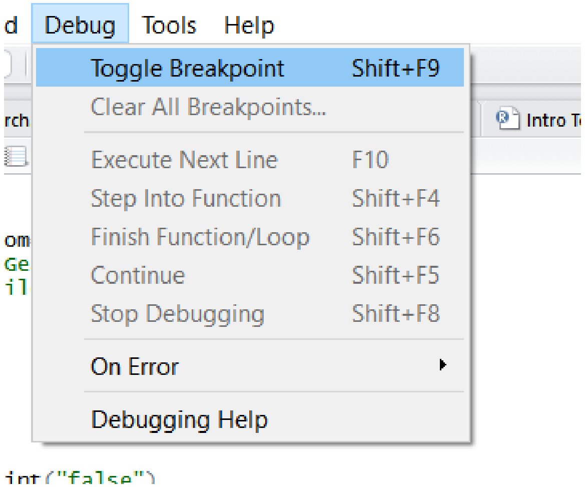

Run Selected Line(s) Ctrl+Enter
Re-Run Previous Ctrl+Shift+P
Run Region
Source Ctrl+Shift+S
Source with Echo Ctrl+Shift+Enter
Source File... Ctrl+Alt+G
Run From Beginning To Line Ctrl+Alt+B
Run From Line to End Ctrl+Alt+E
Run Function Definition Ctrl+Alt+F
Run Code Section Ctrl+Alt+T
Run All Ctrl+Alt+R





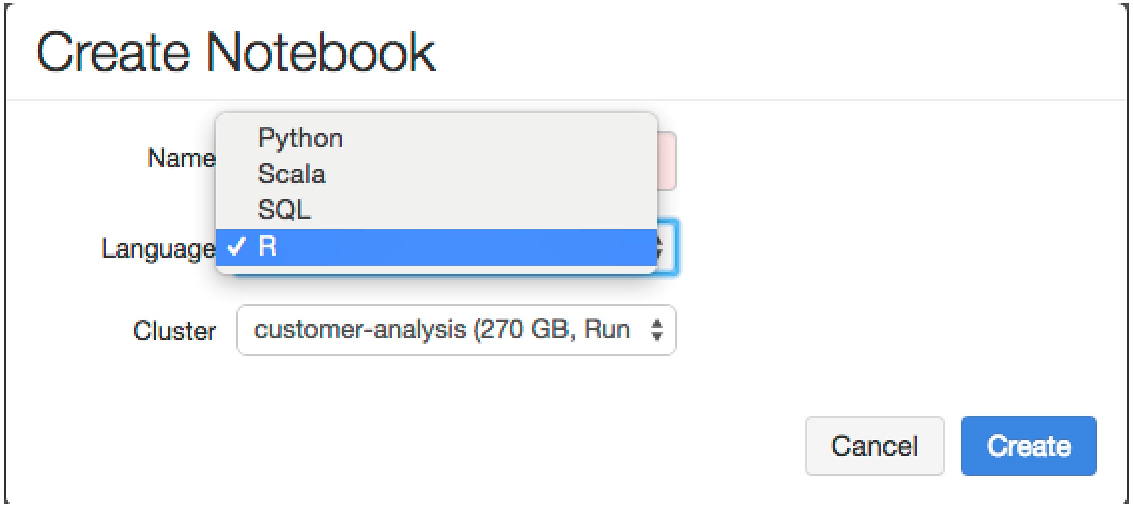
RStudio Debugging

- Breakpoints (Shift+F9)
- R Functions
  - browser()
  - debugonce()
- Environment Pane
  - Traceback(Callstack)
- Console
  - Step into function (Shift+F4)
  - Finish Function (Shift+F6)
  - Continue Running (Shift+F5)
  - Stop Debugging (Shift+F8)



Azure Databricks

- Azure Databricks
  - R Integration
  - Python
  - Scala
  - Spark SQL



Create Notebook

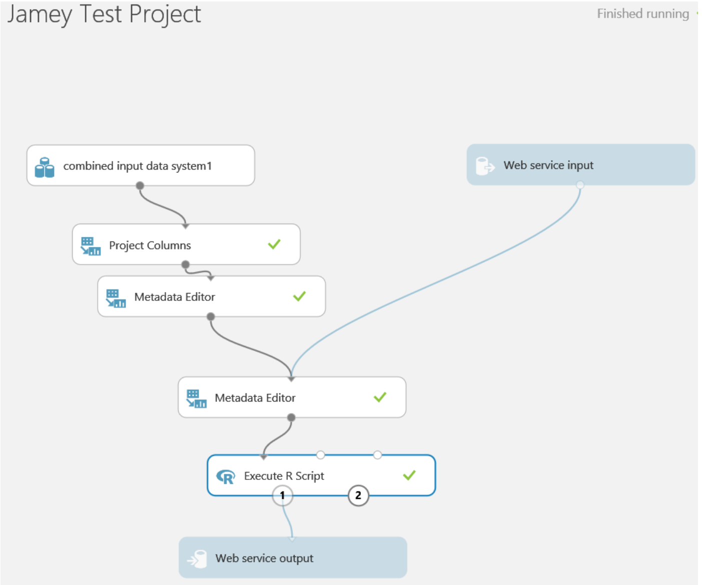
Name

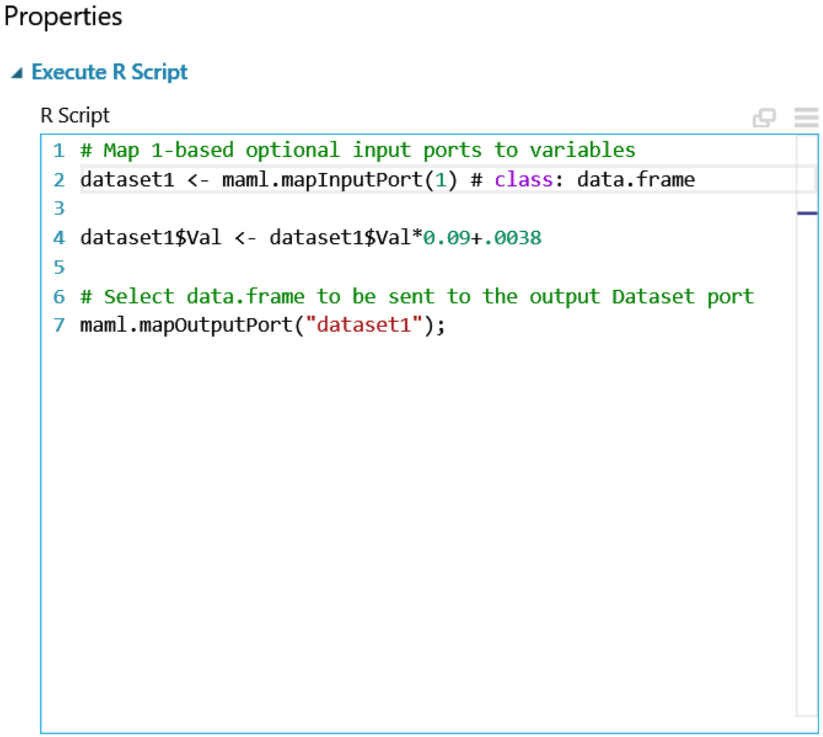
Language ✓ R

Cluster customer-analysis (270 GB, Run

Azure ML

- Azure Machine Learning
  - R Integration

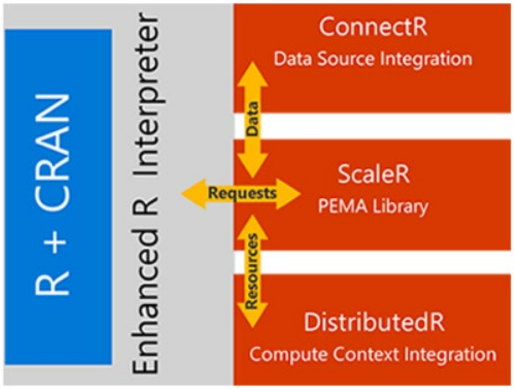




```r
# Map 1-based optional input ports to variables
dataset1 <- maml.mapInputPort(1) # class: data.frame

dataset1$Val <- dataset1$Val*0.09+.0038

# Select data.frame to be sent to the output Dataset port
maml.mapOutputPort("dataset1");
```

MS Machine Learning Services

- Enterprise Class R, Python and Java (2019)
- Built on Revolution Analytics acquistion
- SQL Server 2016 R Support via R Server
  - https://www.microsoft.com/en-us/server-cloud/products/r-server/



Source: Microsoft Website (URL above)

SQL 2016+ and R

- Leverages the MS R Server
- https://docs.microsoft.com/en-us/sql/advanced-analytics/what-is-sql-server-machine-learning?view=sql-server-2017

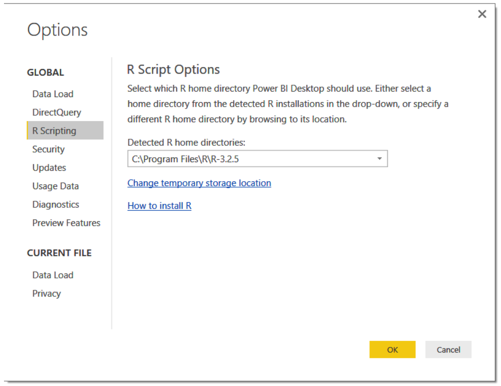
SQL 2016+ and R

- **SQL Server R Services Tutorials**
  - https://msdn.microsoft.com/en-US/library/mt591993.aspx
- DEMO - iris-sepal-example.sql
  - **sp_execute_external_script (Transact-SQL)**
  - https://msdn.microsoft.com/en-us/library/mt604368.aspx

```
sp_execute_external_script
    @language = N'language' ,
    @script = N'script',
    @input_data_1 = ] 'input_data_1'
    [ , @input_data_1_name = ] N'input_data_1_name' ]
    [ , @output_data_1_name = 'output_data_1_name' ]
    [ WITH <execute_option> [ ,...n ] ]
[;]
```

Power BI

- **Running R Scripts in Power BI Desktop**
  - https://powerbi.microsoft.com/en-us/documentation/powerbi-desktop-r-scripts/
  - https://powerbi.microsoft.com/en-us/blog/announcing-preview-of-r-visuals-in-power-bi-desktop/
- **Demo – mtcars.pbix**



Options

GLOBAL
- Data Load
- DirectQuery
- R Scripting
- Security
- Updates
- Usage Data
- Diagnostics
- Preview Features

CURRENT FILE
- Data Load
- Privacy

R Script Options

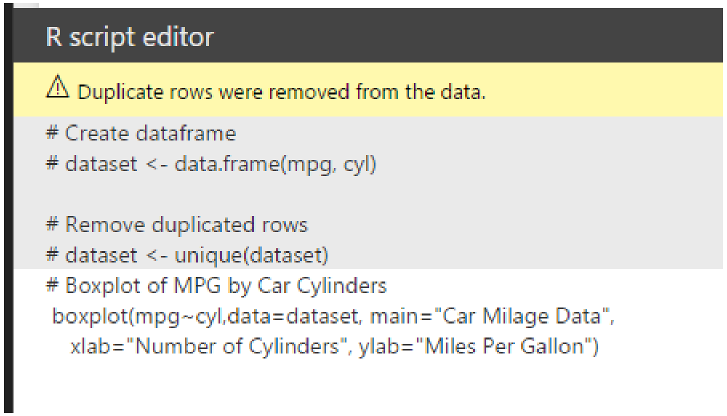
Select which R home directory Power BI Desktop should use. Either select a home directory from the detected R installations in the drop-down, or specify a different R home directory by browsing to its location.

Detected R home directories: C:\Program Files\R\R-3.2.5

Change temporary storage location

How to install R

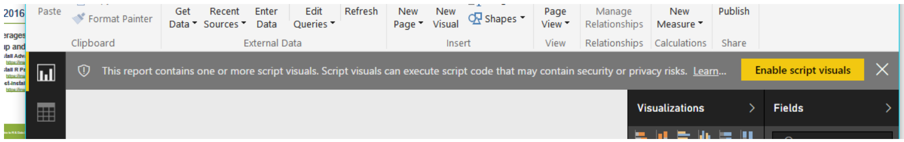
Options Needed



R script editor

⚠ Duplicate rows were removed from the data.

```
# Create dataframe
# dataset <- data.frame(mpg, cyl)

# Remove duplicated rows
# dataset <- unique(dataset)
# Boxplot of MPG by Car Cylinders
 boxplot(mpg~cyl,data=dataset, main="Car Milage Data",
   xlab="Number of Cylinders", ylab="Miles Per Gallon")
```



This report contains one or more script visuals. Script visuals can execute script code that may contain security or privacy risks. Learn...

Resources

- UCLA idre
  - http://statistics.ats.ucla.edu/stat/r/
- R-Bloggers (sign up for daily email)
  - http://www.r-bloggers.com/
- Quick-R
  - http://www.statmethods.net/
  - R in Action (book to go with website)
- Hadley Wickham
  - http://hadley.nz/

Thank You Sponsors!

Visit the Sponsor tables to enter their end of day raffles.

Turn in your completed Event Evaluation form at the end of the day in the Registration area to be entered in additional drawings.

Questions?

Thank you for attending!

- @STATCowboy
- http://STATCowboy.com
- https://github.com/STATCowboy/CodeLIkePirate
  - Download Demos and PPT