Want to follow along with this session using R? Download the script and data from the session scheduler. Also download R and RStudio.

It's easy to follow along!




YOU CAN
CHOOSE
TO SEE
ANALYTICS
DIFFERENTLY.
RED
PILL
ANALYTICS
RedPillAnalytics.com

Using R for Data Profiling

Michelle Kolbe

medium.com/@datacheesehead

@mekolbe

linkedin.com/in/michellekolbe

michelle.kolbe@redpillanalytics.com

© 2016 RED PILL Analytics

© 2016 RED PILL Analytics



Yes! Gartner estimated “more than 25 percent of critical data within Fortune 1000 enterprises” to be flawed.

TDWI stated that “data quality problems cost US businesses more than $600 billion a year” and poor data quality leads to failure and delays of many high profile IT projects.

Lack of trust in the data results in reduced or discontinued BI usage

Source: https://datasourceconsulting.com/data-profiling/

What to Check for?

- Accuracy
- Consistency
- Completeness
- Uniqueness
- Distribution
- Range

© 2016 RED PILL Analytics


© 2016 RED PILL Analytics
Why Profile Your Data?

Benefits

- Trust in data
- Find problems in advance
- Shorten development time on projects
- Improve understanding of data & business knowledge

© 2016 RED PILL Analytics

© 2016 RED PILL Analytics



Why R?

Why R?

- Free!
- Easy to use
- Flexible
- Powerful analytics
- Great community!

© 2016 RED PILL Analytics

Flexible because it's a language
And you can use varied datasets, do data manipulation, & run stats models

© 2016 RED PILL Analytics



Getting Started in R

What is R?



- A programming environment
- Fairly simple to use & understand
- Allows a user to manipulate & analyze data
- Open source
- Real power comes from available packages you can install from LARGE community
- Easy to learn with programming background
- Con: Memory management & speed vs C++ or Python

© 2016 RED PILL Analytics

Tools for R

- First download R from r-project.org
- Then download R Studio, the best R IDE



 © 2016 RED PILL Analytics

R Basics

- Case sensitive
- <- assigns to a variable
- # begins a comment
- ??<keyword> will search R documentation for help



© 2016 RED PILL Analytics

Using Packages

- First install
  `install.packages("<package name>")`
- Once installed, load the package
  `library("<package name>")`
- Note that every time you open R you'll need to load the packages you'll be using
- You'll see your packages that are installed and loaded in R Studio



 © 2016 RED PILL Analytics

Connecting to Data in R

- Data should be read into R and stored into an object
- Easiest with CSV
- Can download datasets from a url or located on a drive

```
d <- read.csv("http://www.ats.ucla.edu/stat/data/hsb2.csv")
```



© 2016 RED PILL Analytics

Connecting to Oracle

- RODBC
  - Load package in R
    ```
    library(RODBC)
    ```
  - View available data sources
    ```
    odbcDataSources()
    ```
  - Can read tables and send sql queries
    ```
    con <- odbcConnect("Oracle Sample", uid="system", pwd="oracle")
    d <- sqlQuery(con, "select sysdate from dual")
    ```



© 2016 RED PILL Analytics

Connecting to Oracle

- RJDBC
  - Load Package
    ```
    library(RJDBC)
    ```
  - Create connection driver
    ```
    jdbcDriver <- JDBC(driverClass="oracle.jdbc.OracleDriver",
    classPath="lib/ojdbc6.jar")
    ```
  - Open Connection
    ```
    jdbcConnection <- dbConnect(jdbcDriver, "jdbc:oracle:thin@//
    database.hostname.com:port/service_name_or_sid", "username",
    "password")
    ```
  - Query
    ```
    dbGetQuery(jdbcConnection, "select sysdate from dual")
    ```
  - Close Connection
    ```
    dbDisconnect(jdbcConnection)
    ```



 © 2016 RED PILL Analytics

Source for RJDBC: http://www.r-bloggers.com/connecting-r-to-an-oracle-database-with-rjdbc/

ROracle

- Open Source but maintained by Oracle
- Faster: 79 times faster than RJDBC and 2.5 times faster than RODBC
- Provides scalability and stability



© 2016 RED PILL Analytics

In full disclosure, I haven’t used this yet but it appears to still require you to pull the data into R into a data frame

Variables

- Can store data in variables using <- or =
- Do not need to define variable first
- RStudio shows your variables on the right

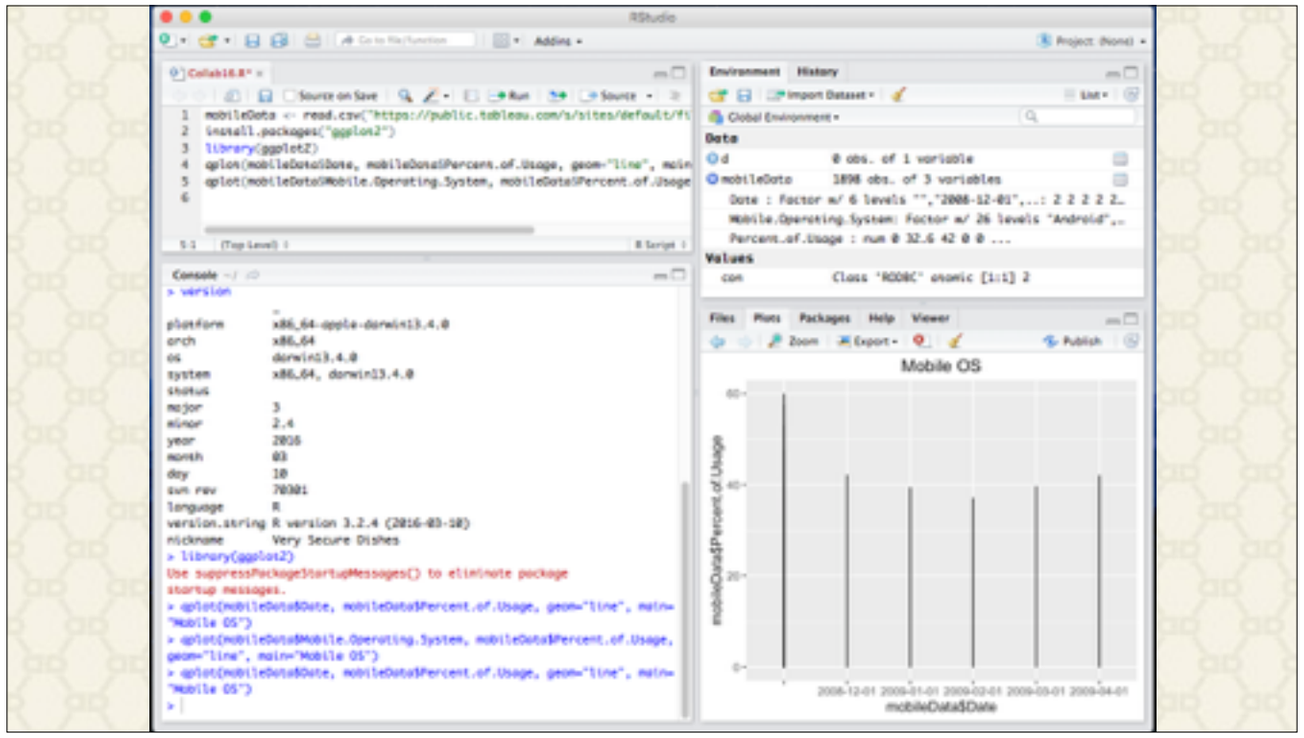


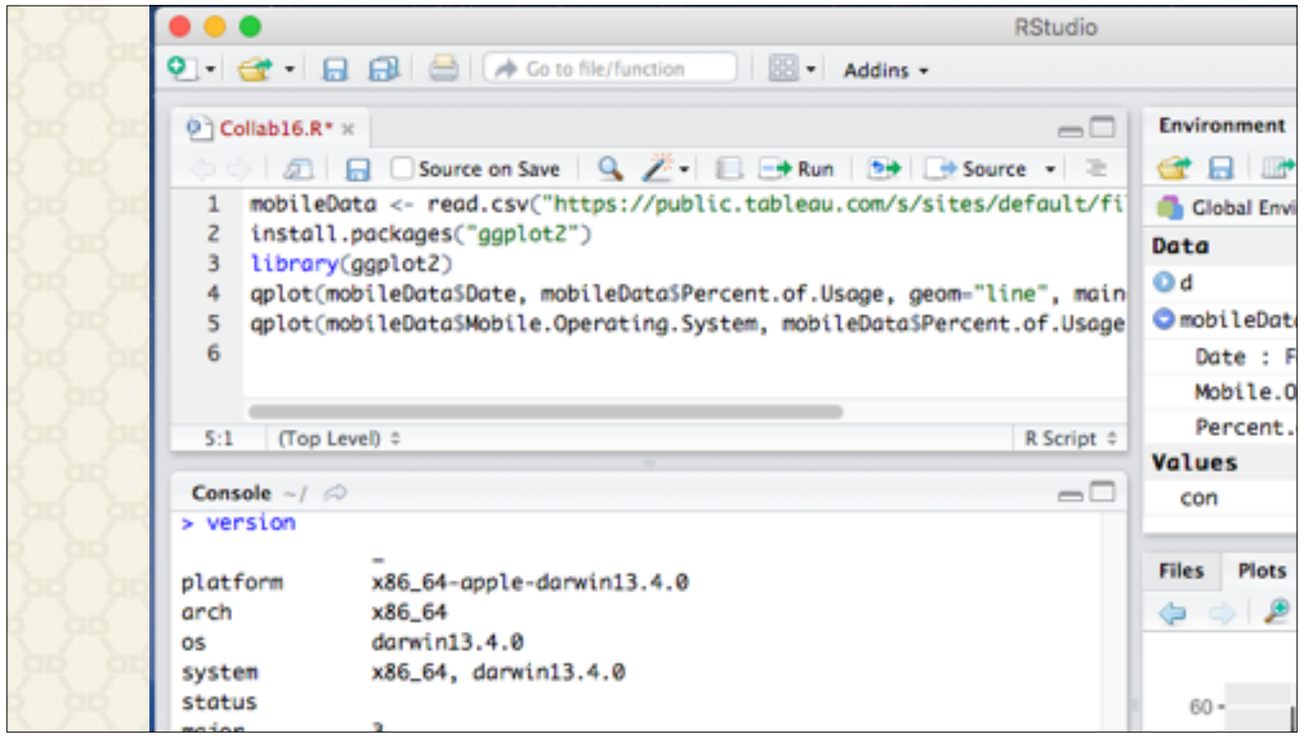
© 2016 RED PILL Analytics

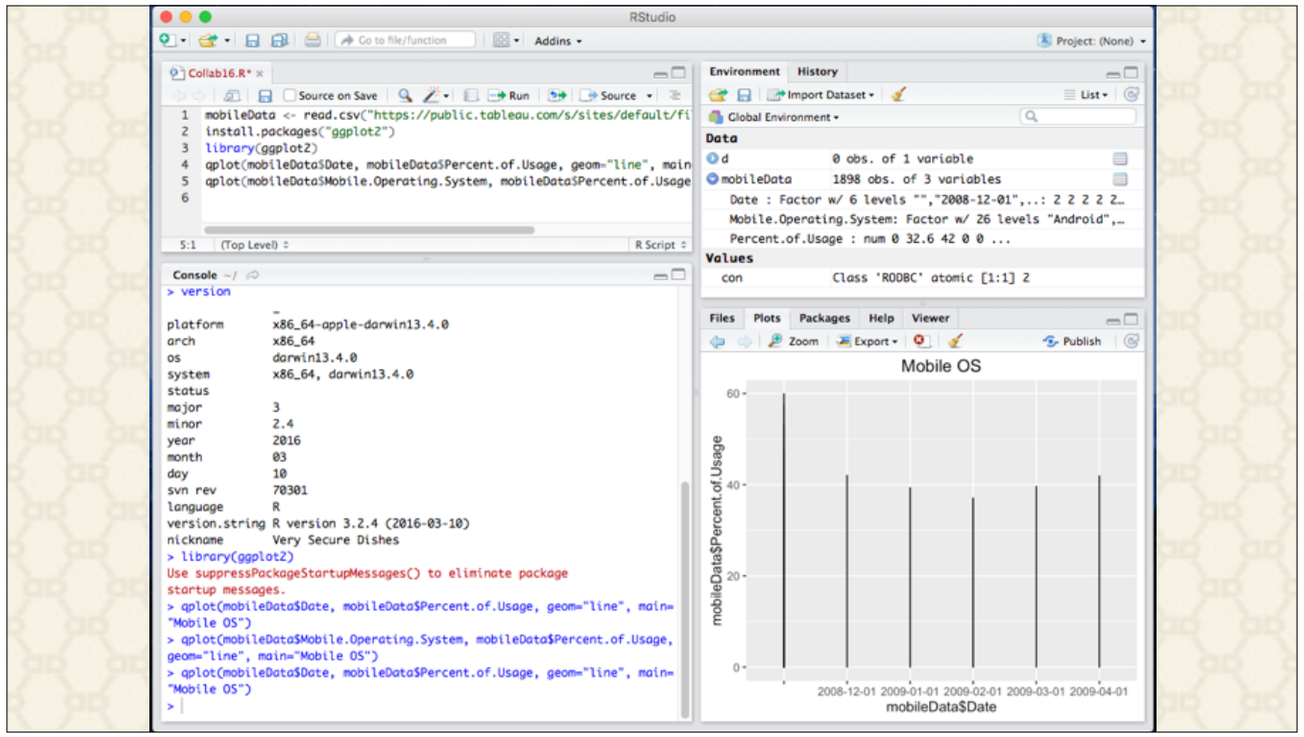
© 2016 RED PILL Analytics

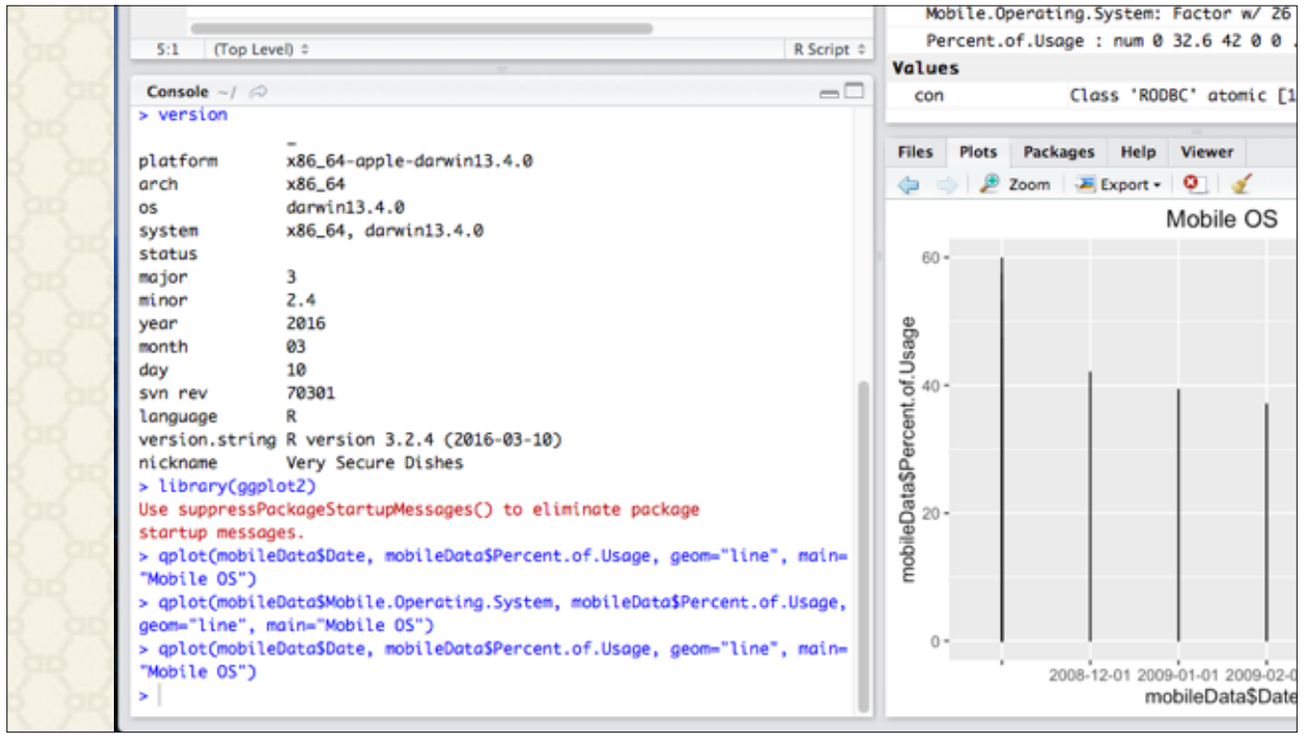


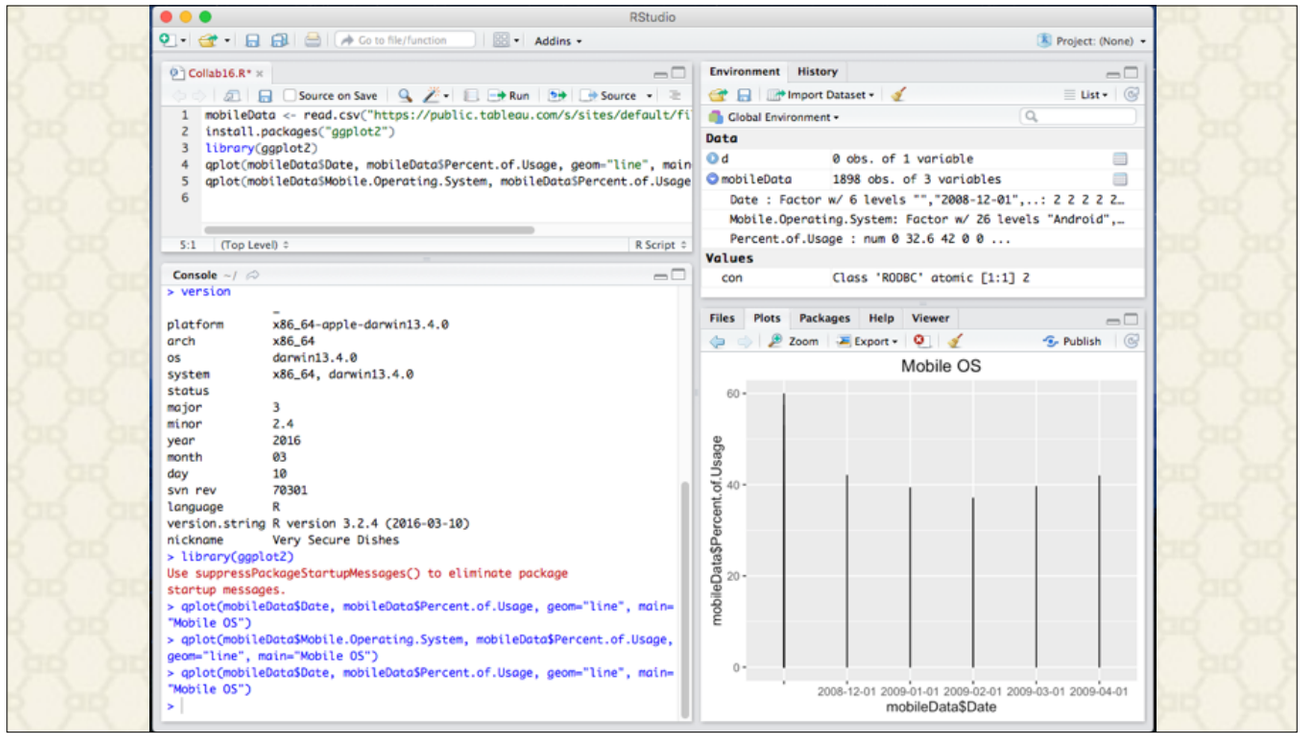
Using R Studio

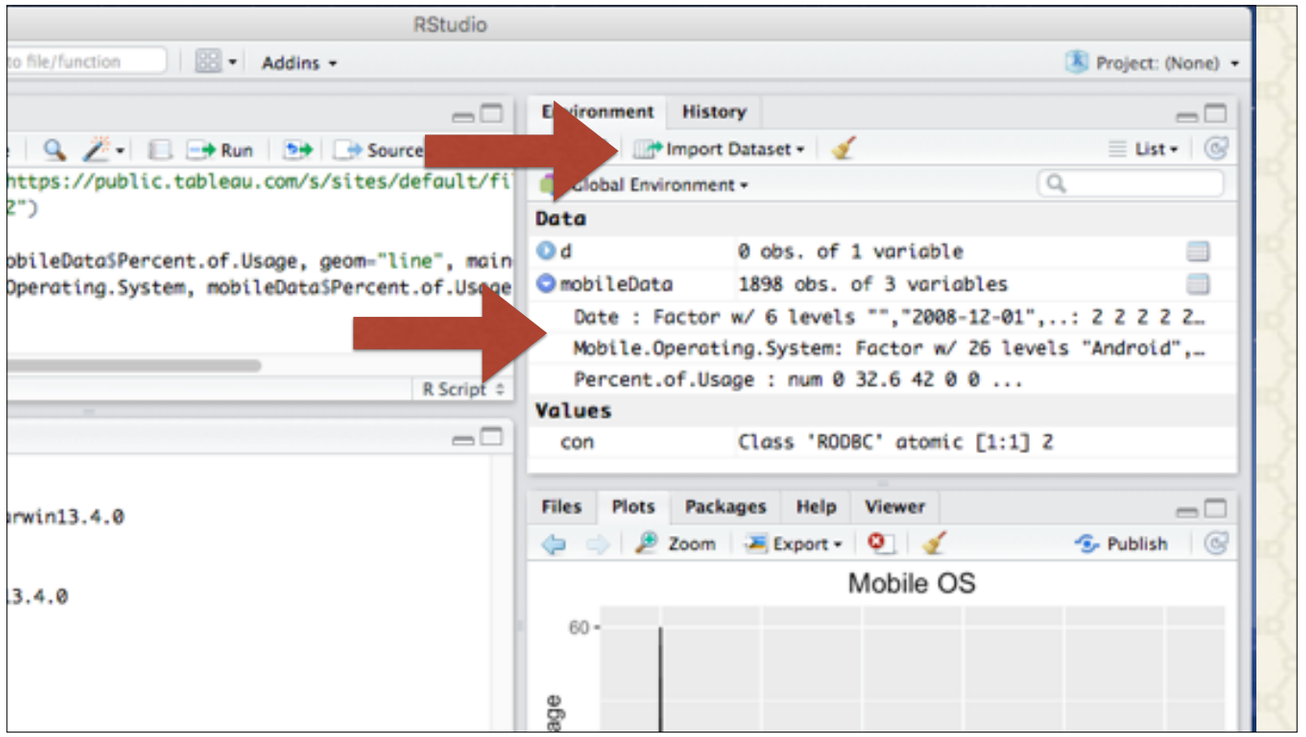

RStudio
mobileData <- read.csv("https://public.tableau.com/s/sites/default/fi
install.packages("ggplot2")
library(ggplot2)
qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main
qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage
Environment
History
Import Dataset
Global Environment
Data
d 0 obs. of 1 variable
mobileData 1898 obs. of 3 variables
Date : Factor w/ 6 levels "","2008-12-01",...: 2 2 2 2 2..
Mobile.Operating.System: Factor w/ 26 levels "Android",..
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con Class 'RODBC' onomic [1:1] 2
Console
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
minor 2.4
year 2016
month 03
day 10
svn rev 70301
language R
version.string R version 3.2.4 (2016-03-10)
nickname Very Secure Dishes
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package
startup messages.
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
> qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage,
geom="line", main="Mobile OS")
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
60
40
20
0
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date

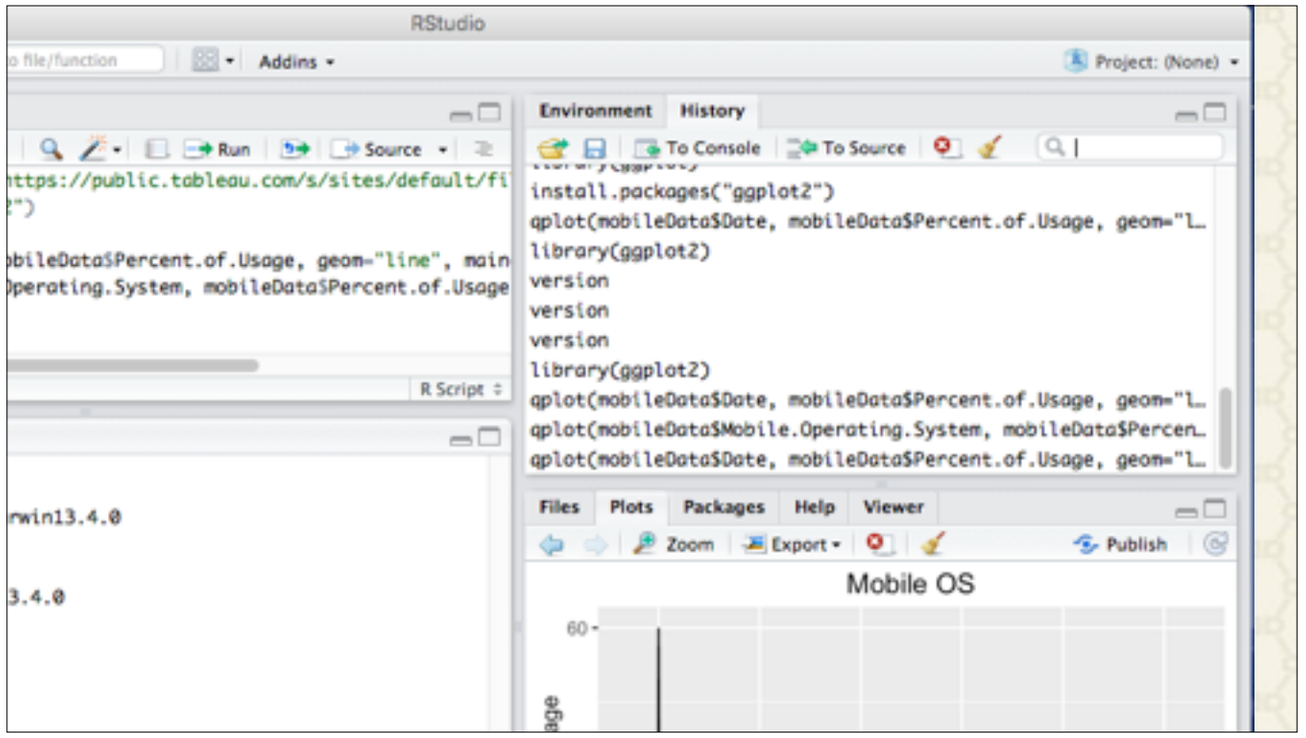

RStudio
Go to file/function
Addins
Collab16.R*
Source on Save
Run
Source
1 mobileData <- read.csv("https://public.tableau.com/s/sites/default/fi
2 install.packages("ggplot2")
3 library(ggplot2)
4 qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main
5 qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage
6
5:1
(Top Level)
R Script
Console ~/
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
Environment
Global Envi
Data
d
mobileDat
Date : F
Mobile.O
Percent.
Values
con
Files
Plots
60

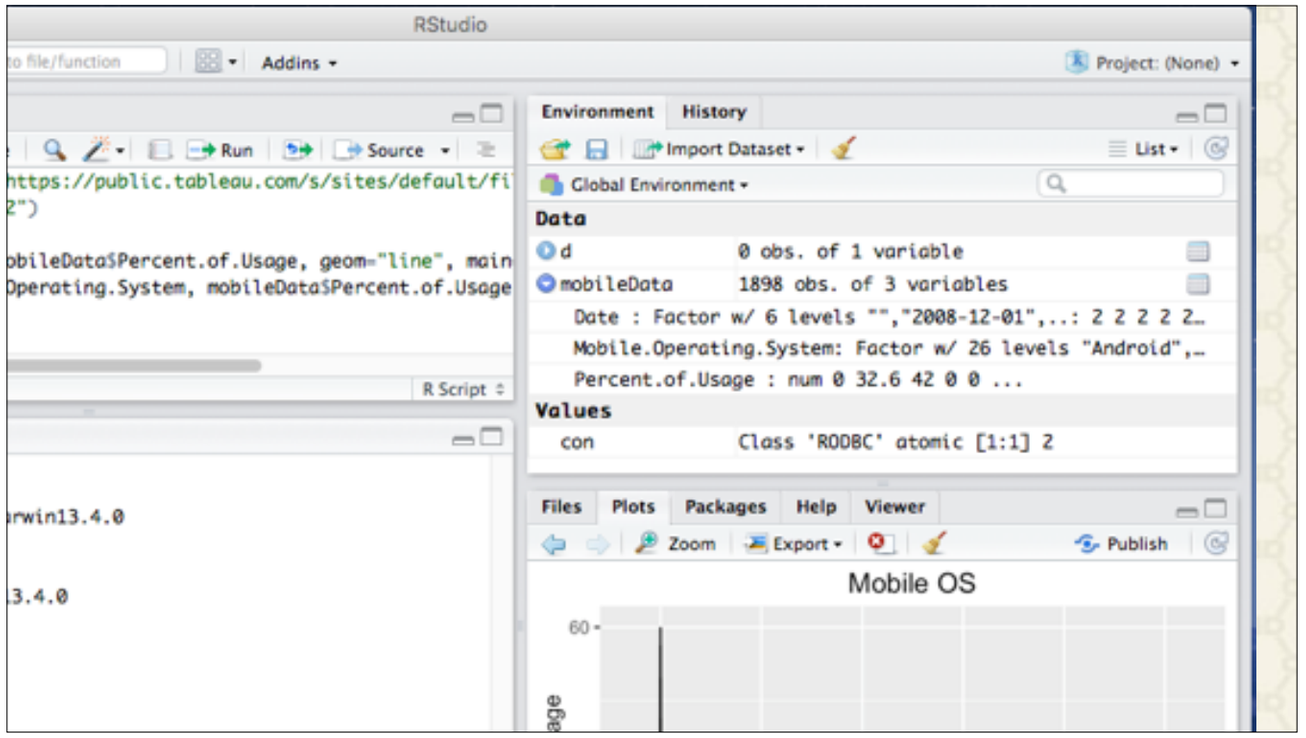

RStudio
Addins
Project: (None)
Collab16.R*
Source on Save
Run
Source
1 mobileData <- read.csv("https://public.tableau.com/s/sites/default/fi
2 install.packages("ggplot2")
3 library(ggplot2)
4 qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main
5 qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage
6
5:1
(Top Level)
R Script
Console ~/
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
minor 2.4
year 2016
month 03
day 10
svn rev 70301
language R
version.string R version 3.2.4 (2016-03-10)
nickname Very Secure Dishes
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package
startup messages.
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
> qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage,
geom="line", main="Mobile OS")
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
>
Environment
History
Import Dataset
List
Global Environment
Data
d 0 obs. of 1 variable
mobileData 1898 obs. of 3 variables
Date : Factor w/ 6 levels "","2008-12-01",..: 2 2 2 2 2…
Mobile.Operating.System: Factor w/ 26 levels "Android",…
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con Class 'RODBC' atomic [1:1] 2
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
0
20
40
60
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date

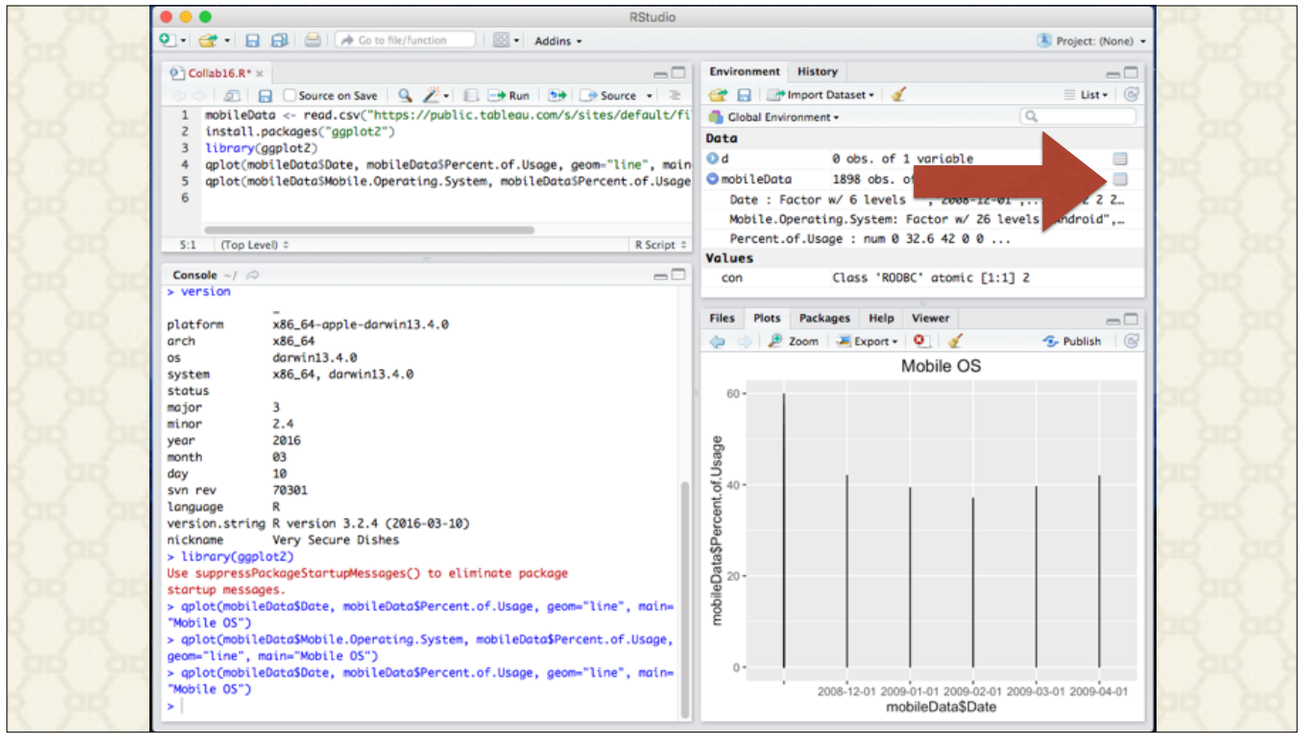

5:1
(Top Level)
R Script
Console ~/
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
minor 2.4
year 2016
month 03
day 10
svn rev 70301
language R
version.string R version 3.2.4 (2016-03-10)
nickname Very Secure Dishes
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package
startup messages.
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main="Mobile OS")
> qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage, geom="line", main="Mobile OS")
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main="Mobile OS")
>
Mobile.Operating.System: Factor w/ 26
Percent.of.Usage : num 0 32.6 42 0 0
Values
con Class 'RODBC' atomic [1
Files
Plots
Packages
Help
Viewer
Zoom
Export
Mobile OS
mobileData$Percent.of.Usage
60
40
20
0
2008-12-01
2009-01-01
mobileData$Date

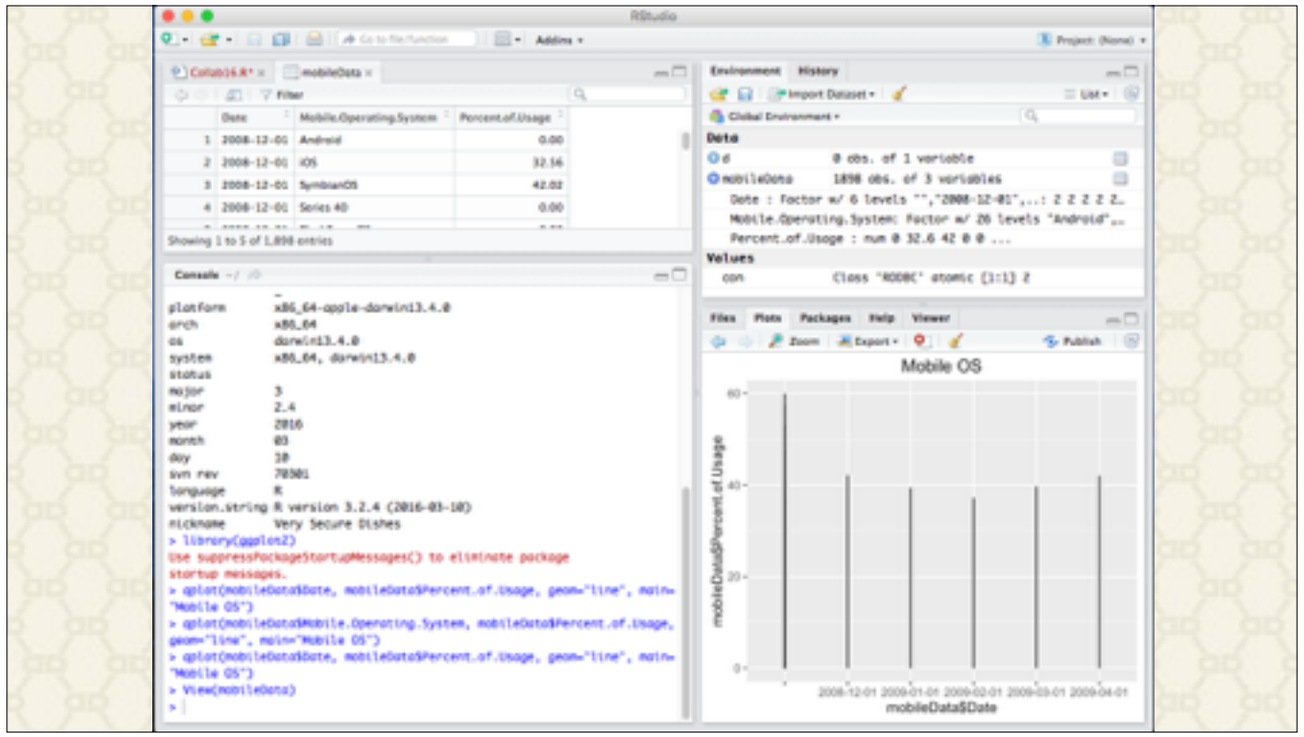

RStudio
Addins
Project: (None)
Collab16.R*
Source on Save
Run
Source
1 mobileData <- read.csv("https://public.tableau.com/s/sites/default/fi
2 install.packages("ggplot2")
3 library(ggplot2)
4 qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main
5 qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage
6
5:1
(Top Level)
R Script
Console ~/
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
minor 2.4
year 2016
month 03
day 10
svn rev 70301
language R
version.string R version 3.2.4 (2016-03-10)
nickname Very Secure Dishes
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package
startup messages.
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
> qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage,
geom="line", main="Mobile OS")
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
>
Environment
History
Import Dataset
List
Global Environment
Data
d 0 obs. of 1 variable
mobileData 1898 obs. of 3 variables
Date : Factor w/ 6 levels "","2008-12-01",..: 2 2 2 2 2…
Mobile.Operating.System: Factor w/ 26 levels "Android",…
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con Class 'RODBC' atomic [1:1] 2
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
0
20
40
60
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date

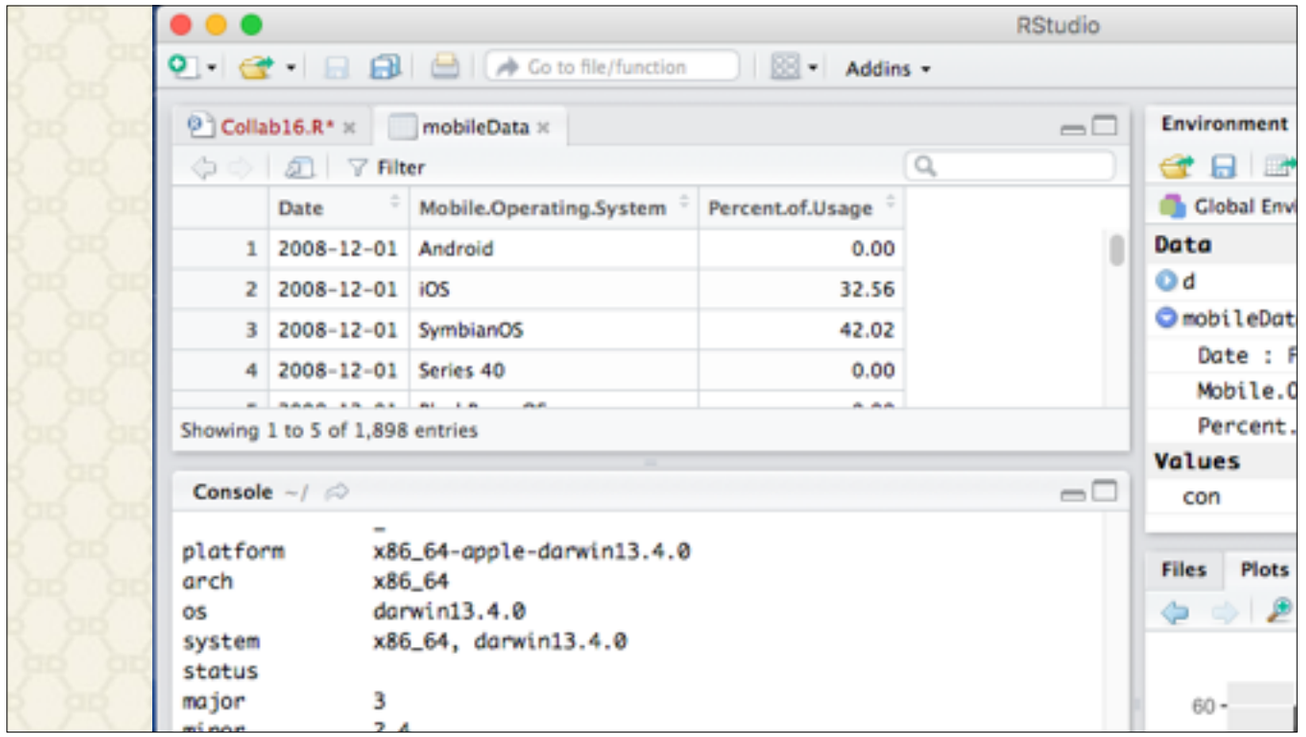

RStudio
Addins
Project: (None)
Run
Source
https://public.tableau.com/s/sites/default/fi
2")
obileData$Percent.of.Usage, geom="line", main
Operating.System, mobileData$Percent.of.Usage
R Script
rwin13.4.0
3.4.0
Environment
History
Import Dataset
List
Global Environment
Data
d
0 obs. of 1 variable
mobileData
1898 obs. of 3 variables
Date : Factor w/ 6 levels "","2008-12-01",..: 2 2 2 2 2…
Mobile.Operating.System: Factor w/ 26 levels "Android",…
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con
Class 'RODBC' atomic [1:1] 2
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
60
age

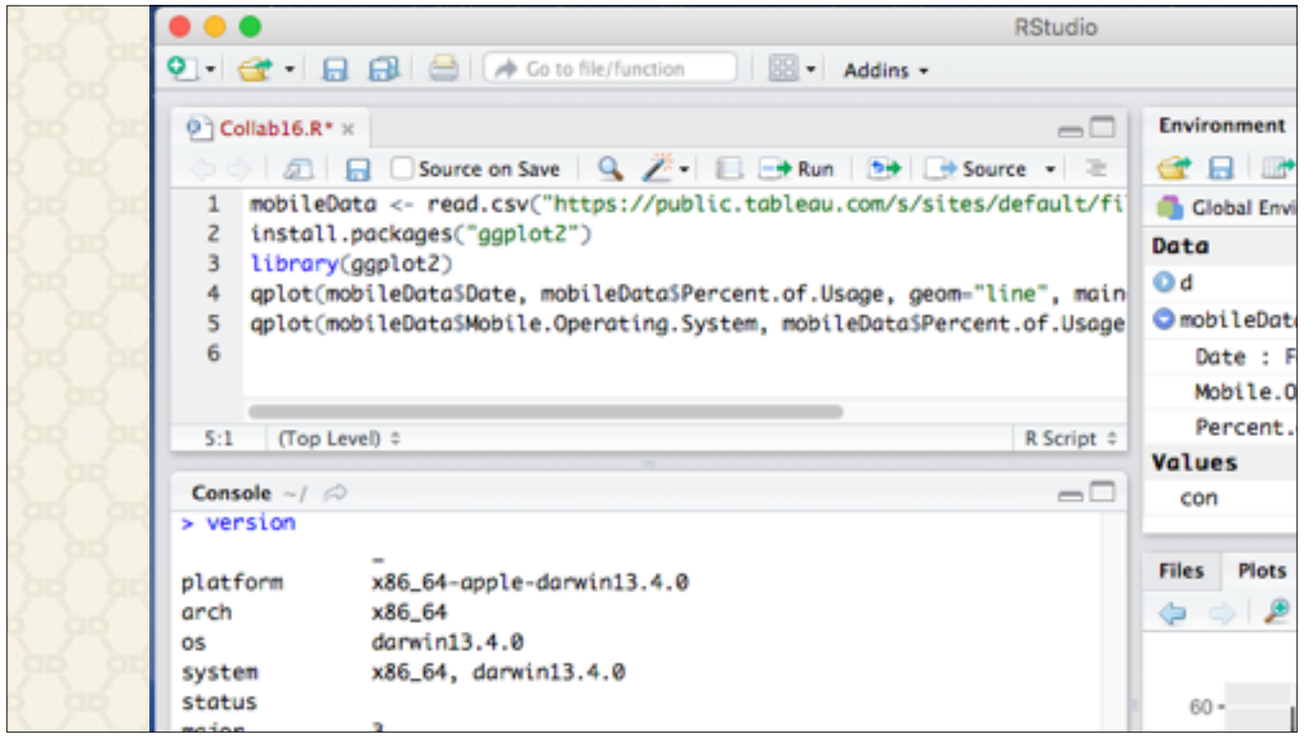

RStudio
Addins
Project: (None)
Run
Source
https://public.tableau.com/s/sites/default/fi
")
obileData$Percent.of.Usage, geom="line", main
Operating.System, mobileData$Percent.of.Usage
R Script
rwin13.4.0
3.4.0
Environment
History
To Console
To Source
install.packages("ggplot2")
qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="l…
library(ggplot2)
version
version
version
library(ggplot2)
qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="l…
qplot(mobileData$Mobile.Operating.System, mobileData$Percen…
qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="l…
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
60
age

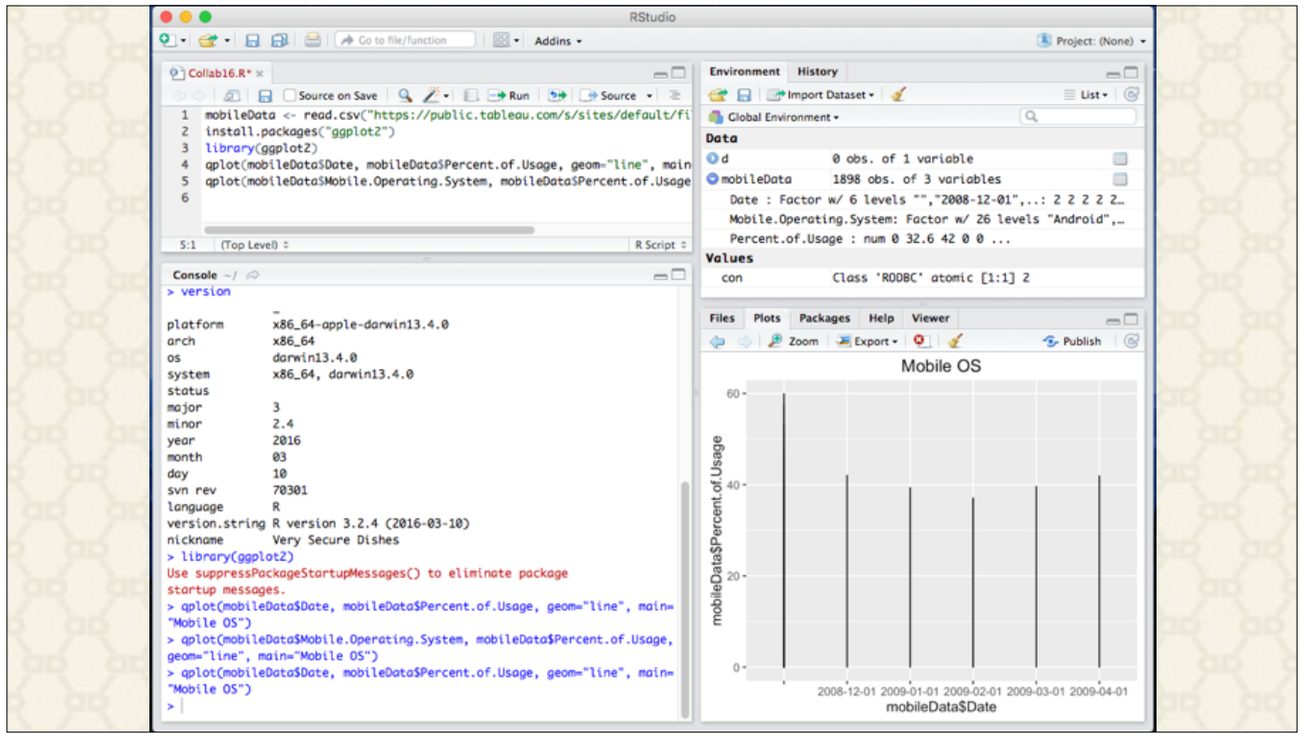

RStudio
Addins
Project: (None)
Run
Source
https://public.tableau.com/s/sites/default/fi
2")
obileData$Percent.of.Usage, geom="line", main
Operating.System, mobileData$Percent.of.Usage
R Script
rwin13.4.0
3.4.0
Environment
History
Import Dataset
List
Global Environment
Data
d
0 obs. of 1 variable
mobileData
1898 obs. of 3 variables
Date : Factor w/ 6 levels "","2008-12-01",..: 2 2 2 2 2…
Mobile.Operating.System: Factor w/ 26 levels "Android",…
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con
Class 'RODBC' atomic [1:1] 2
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
60
age


RStudio
Addins
Project: (None)
Collab16.R*
Source on Save
Run
Source
1 mobileData <- read.csv("https://public.tableau.com/s/sites/default/fi
2 install.packages("ggplot2")
3 library(ggplot2)
4 qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main
5 qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage
6
5:1
(Top Level)
R Script
Console ~/
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
minor 2.4
year 2016
month 03
day 10
svn rev 70301
language R
version.string R version 3.2.4 (2016-03-10)
nickname Very Secure Dishes
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package
startup messages.
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
> qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage,
geom="line", main="Mobile OS")
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
>
Environment
History
Import Dataset
List
Global Environment
Data
d 0 obs. of 1 variable
mobileData 1898 obs. of
Date : Factor w/ 6 levels
Mobile.Operating.System: Factor w/ 26 levels
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con Class 'RODBC' atomic [1:1] 2
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
0
20
40
60
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date

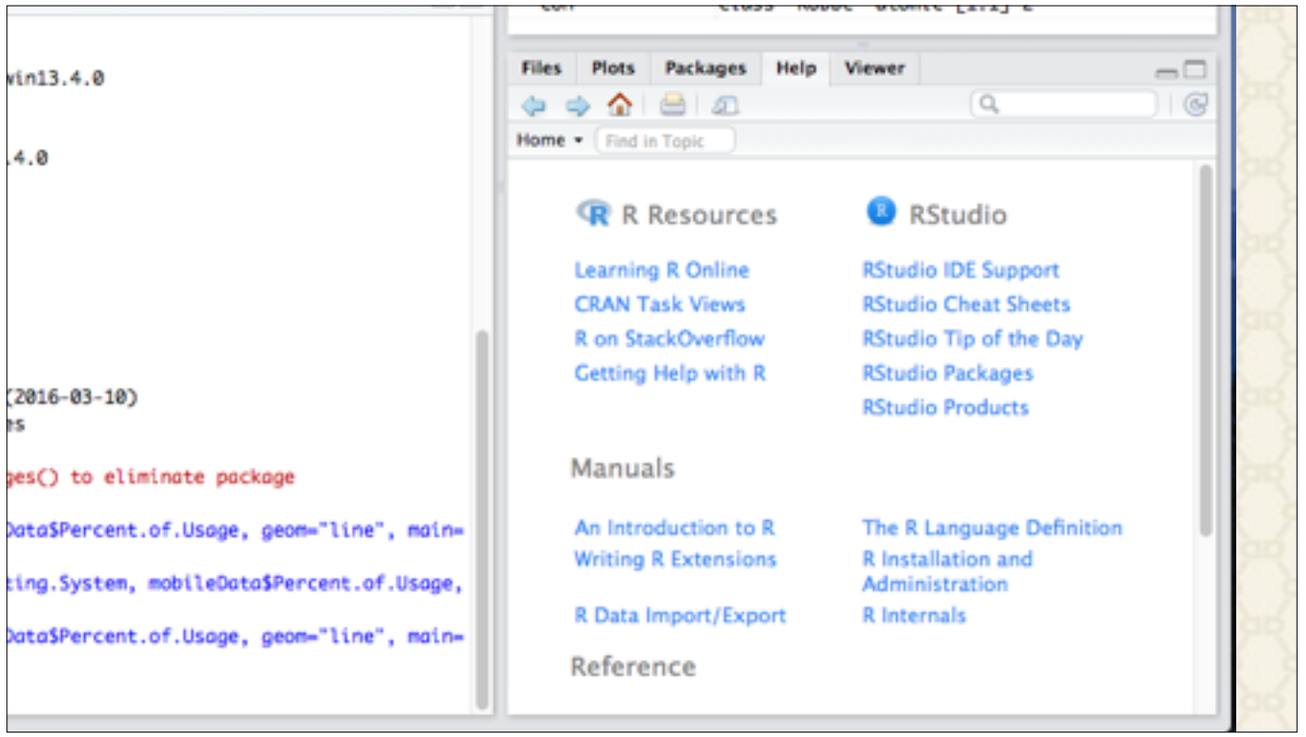

RStudio
Addins
Project: (None)
Coltab16.R*
mobileData
Filter
Date
Mobile.Operating.System
Percent.of.Usage
1 2008-12-01 Android 0.00
2 2008-12-01 iOS 32.16
3 2008-12-01 SymbianOS 42.02
4 2008-12-01 Series 40 0.00
Showing 1 to 5 of 1,898 entries
Console ~/
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
minor 2.4
year 2016
month 03
day 10
svn rev 70301
language R
version.string R version 3.2.4 (2016-03-10)
nickname Very Secure Dishes
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package
startup messages.
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main="Mobile OS")
> qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage, geom="line", main="Mobile OS")
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main="Mobile OS")
> View(mobileData)
>
Environment
History
Import Dataset
List
Global Environment
Data
d 0 obs. of 1 variable
mobileData 1898 obs. of 3 variables
Date : Factor w/ 6 levels "","2008-12-01",...: 2 2 2 2 2 2 ...
Mobile.Operating.System: Factor w/ 20 levels "Android",...
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con Class 'RODBC' atomic [1:1] 2
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
0
20
40
60
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date

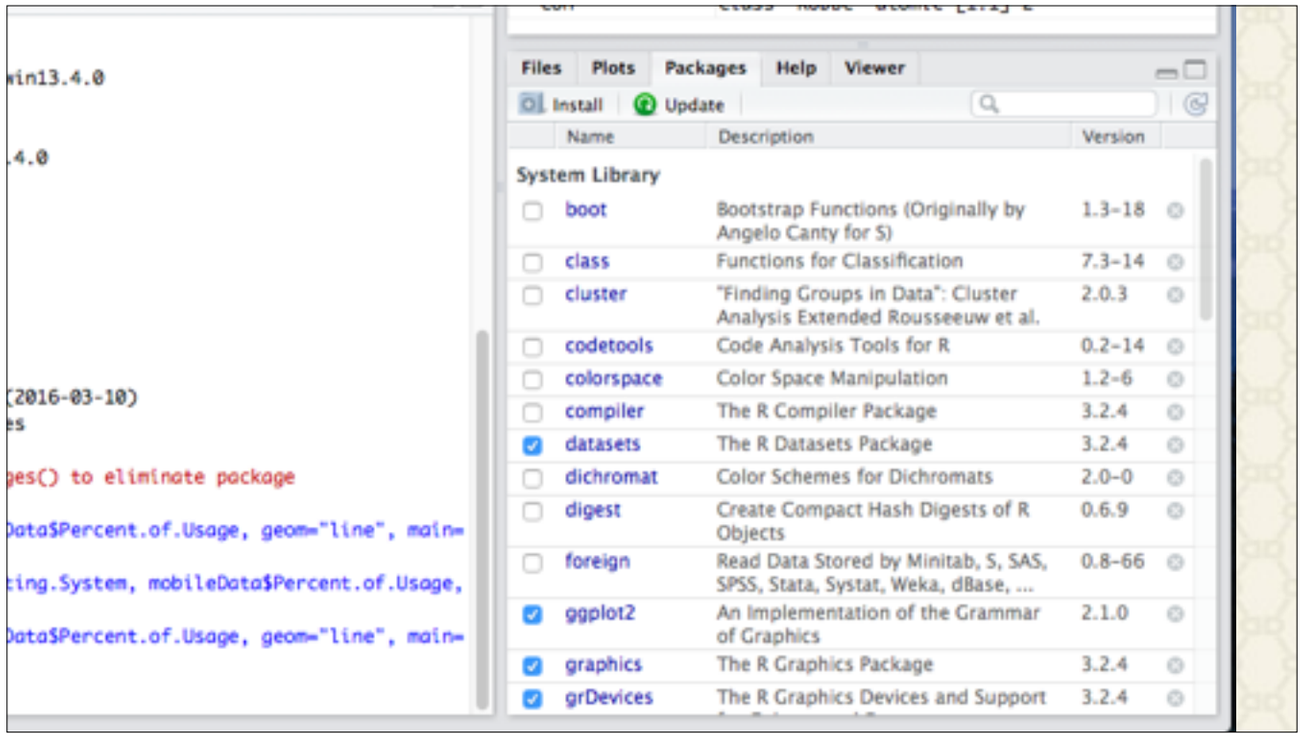

RStudio
Go to file/function
Addins
Collab16.R*
mobileData
Filter
Date
Mobile.Operating.System
Percent.of.Usage
1 2008-12-01 Android 0.00
2 2008-12-01 iOS 32.56
3 2008-12-01 SymbianOS 42.02
4 2008-12-01 Series 40 0.00
Showing 1 to 5 of 1,898 entries
Console ~/
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
Environment
Global Env
Data
d
Values
con
Files
Plots
60


RStudio
Go to file/function
Addins
Collab16.R*
Source on Save
Run
Source
1 mobileData <- read.csv("https://public.tableau.com/s/sites/default/fi
2 install.packages("ggplot2")
3 library(ggplot2)
4 qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main
5 qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage
6
5:1
(Top Level)
R Script
Console ~/
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
Environment
Global Envi
Data
d
mobileDat
Date : F
Mobile.O
Percent.
Values
con
Files
Plots
60

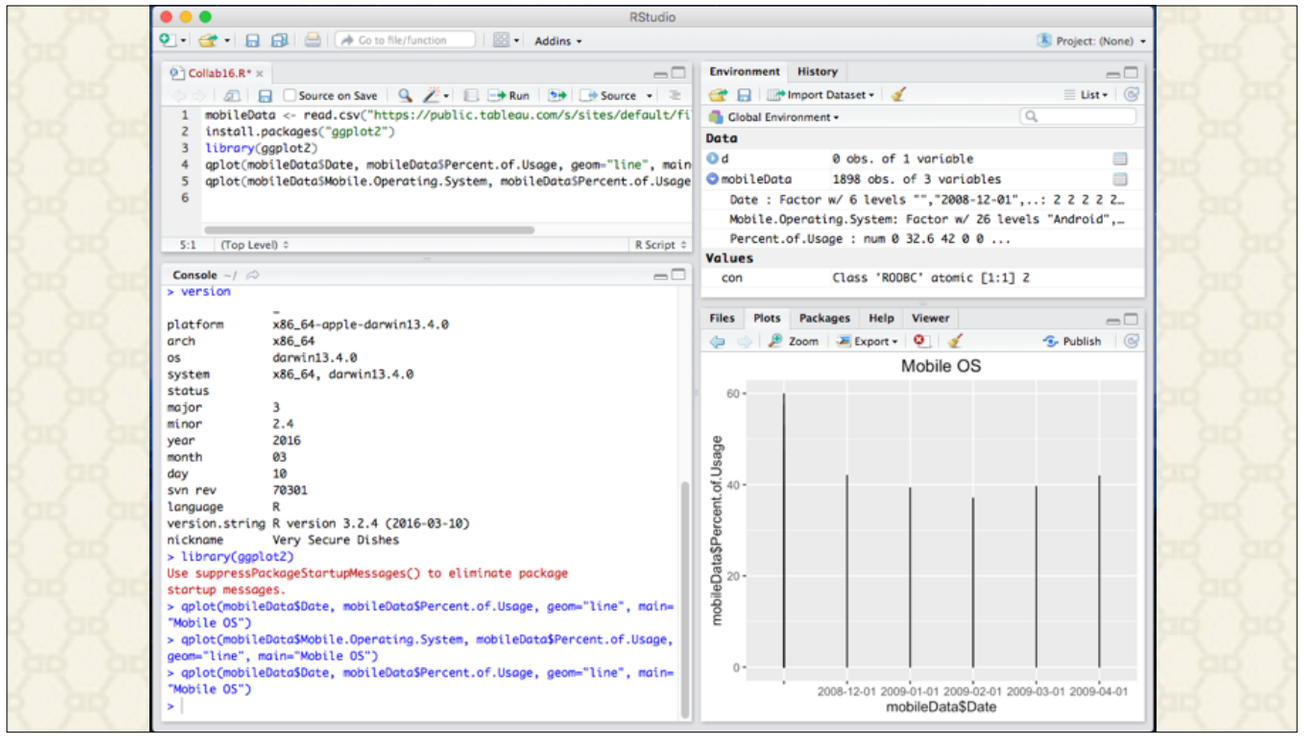

RStudio
Addins
Project: (None)
Collab16.R*
Source on Save
Run
Source
1 mobileData <- read.csv("https://public.tableau.com/s/sites/default/fi
2 install.packages("ggplot2")
3 library(ggplot2)
4 qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main
5 qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage
6
5:1
(Top Level)
R Script
Console ~/
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
minor 2.4
year 2016
month 03
day 10
svn rev 70301
language R
version.string R version 3.2.4 (2016-03-10)
nickname Very Secure Dishes
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package
startup messages.
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
> qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage,
geom="line", main="Mobile OS")
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
>
Environment
History
Import Dataset
List
Global Environment
Data
d 0 obs. of 1 variable
mobileData 1898 obs. of 3 variables
Date : Factor w/ 6 levels "","2008-12-01",..: 2 2 2 2 2…
Mobile.Operating.System: Factor w/ 26 levels "Android",…
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con Class 'RODBC' atomic [1:1] 2
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
0
20
40
60
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date


vin13.4.0
.4.0
(2016-03-10)
es
ges() to eliminate package
Data$Percent.of.Usage, geom="line", main=
ting.System, mobileData$Percent.of.Usage,
Data$Percent.of.Usage, geom="line", main=
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
60
40
20
0
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date


vin13.4.0
.4.0
(2016-03-10)
es
ges() to eliminate package
Data$Percent.of.Usage, geom="line", main=
ting.System, mobileData$Percent.of.Usage,
Data$Percent.of.Usage, geom="line", main=
Files
Plots
Packages
Help
Viewer
Home
Find in Topic
R Resources
Learning R Online
CRAN Task Views
R on StackOverflow
Getting Help with R
RStudio
RStudio IDE Support
RStudio Cheat Sheets
RStudio Tip of the Day
RStudio Packages
RStudio Products
Manuals
An Introduction to R
Writing R Extensions
R Data Import/Export
The R Language Definition
R Installation and Administration
R Internals
Reference


win13.4.0
.4.0
(2016-03-10)
es
ges() to eliminate package
Data$Percent.of.Usage, geom="line", main=
ting.System, mobileData$Percent.of.Usage,
Data$Percent.of.Usage, geom="line", main=
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
System Library
boot
Bootstrap Functions (Originally by Angelo Canty for S)
1.3-18
class
Functions for Classification
7.3-14
cluster
"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.
2.0.3
codetools
Code Analysis Tools for R
0.2-14
colorspace
Color Space Manipulation
1.2-6
compiler
The R Compiler Package
3.2.4
datasets
The R Datasets Package
3.2.4
dichromat
Color Schemes for Dichromats
2.0-0
digest
Create Compact Hash Digests of R Objects
0.6.9
foreign
Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...
0.8-66
ggplot2
An Implementation of the Grammar of Graphics
2.1.0
graphics
The R Graphics Package
3.2.4
grDevices
The R Graphics Devices and Support
3.2.4


vin13.4.0
.4.0
(2016-03-10)
es
ges() to eliminate package
Data$Percent.of.Usage, geom="line", main=
ting.System, mobileData$Percent.of.Usage,
Data$Percent.of.Usage, geom="line", main=
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
60
40
20
0
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date


RStudio
Addins
Project: (None)
Collab16.R*
Source on Save
Run
Source
1 mobileData <- read.csv("https://public.tableau.com/s/sites/default/fi
2 install.packages("ggplot2")
3 library(ggplot2)
4 qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main
5 qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage
6
5:1
(Top Level)
R Script
Console ~/
> version
platform x86_64-apple-darwin13.4.0
arch x86_64
os darwin13.4.0
system x86_64, darwin13.4.0
status
major 3
minor 2.4
year 2016
month 03
day 10
svn rev 70301
language R
version.string R version 3.2.4 (2016-03-10)
nickname Very Secure Dishes
> library(ggplot2)
Use suppressPackageStartupMessages() to eliminate package
startup messages.
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
> qplot(mobileData$Mobile.Operating.System, mobileData$Percent.of.Usage,
geom="line", main="Mobile OS")
> qplot(mobileData$Date, mobileData$Percent.of.Usage, geom="line", main=
"Mobile OS")
>
Environment
History
Import Dataset
List
Global Environment
Data
d 0 obs. of 1 variable
mobileData 1898 obs. of 3 variables
Date : Factor w/ 6 levels "","2008-12-01",..: 2 2 2 2 2…
Mobile.Operating.System: Factor w/ 26 levels "Android",…
Percent.of.Usage : num 0 32.6 42 0 0 ...
Values
con Class 'RODBC' atomic [1:1] 2
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Mobile OS
mobileData$Percent.of.Usage
0
20
40
60
2008-12-01
2009-01-01
2009-02-01
2009-03-01
2009-04-01
mobileData$Date

© 2016 RED PILL Analytics

Our Data Set to Profile



NFL Offensive player statistics & personal information from 1999-2013

First, Load the Data into R

```
> nfl <- read.csv("~/Downloads/nfl.csv")
>   View(nfl)
```

```
nfl              8198 obs. of 39 variables
 Year : int 2006 2013 2007 2011 2011 2004 2000 2011 2003…
 Player : Factor w/ 2302 levels "A.J. Feeley",..: 1368 1…
 Age : int 27 37 30 28 32 27 27 22 30 29 ...
 Hometown : Factor w/ 952 levels "0","Aberdeen",..: 727 …
 Home.State : Factor w/ 67 levels "","AL","American",..:…
 Tm : Factor w/ 36 levels "2TM","3TM","4TM",..: 30 14 24…
 G : int 16 16 16 15 16 16 14 16 16 14 ...
```

Collab16.R* × nfl ×

Filter

	Year	Player	Age	Hometown	Home.State	Tm	G	GS	C
1	2006	LaDainian Tomlinson	27	Rosebud	TX	SDG	16	16	
2	2013	Peyton Manning	37	New Orleans	LA	DEN	16	16	
3	2007	Tom Brady	30	San Mateo	CA	NWE	16	16	
4	2011	Aaron Rodgers	28	Chico	CA	GNB	15	15	
5	2011	Drew Brees	32	Austin	TX	NOR	16	16	
6	2004	Daunte Culpepper	27	Ocala	FL	MIN	16	16	
7	2000	Marshall Faulk	27	New Orleans	LA	STL	14	14	
8	2011	Cam Newton	22	College Park	GA	CAR	16	16	
9	2003	Priest Holmes	30	Fort Smith	AR	KAN	16	16	
10	2002	Priest Holmes	29	Fort Smith	AR	KAN	14	14	

Showing 1 to 11 of 8,198 entries

© 2016 RED PILL Analytics

Summarize the Data

- Summary is an R function to show you basic details about each column in your dataset

```
> summary(nfl)
```

```
     Year                  Player          Age
Min.   :1999   Steve Smith    :  18   Min.   :21.00
1st Qu.:2002   Alex Smith     :  17   1st Qu.:24.00
Median :2006   Adrian Peterson:  15   Median :26.00
Mean   :2006   Brandon Stokley:  15   Mean   :26.65
3rd Qu.:2010   Matt Hasselbeck:  15   3rd Qu.:29.00
Max.   :2013   Tony Gonzalez  :  15   Max.   :44.00
               (Other)        :8103
       Hometown      Home.State         Tm            G
0          : 226   CA     :1081   TAM    : 279   Min.   : 0.00
Houston    : 143   FL     : 746   GNB    : 272   1st Qu.: 7.00
Miami      : 143   TX     : 717   ARI    : 268   Median :14.00
New Orleans: 136   LA     : 410   DEN    : 267   Mean   :11.23
Los Angeles: 132   OH     : 376   OAK    : 265   3rd Qu.:16.00
Washington :  95   PA     : 371   JAX    : 264   Max.   :16.00
(Other)    :7323   (Other):4497   (Other):6583
      GS              Cmp              Att              Yds
Min.   : 0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  -6
1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:   0
Median : 3.000   Median :  0.00   Median :  0.00   Median :   0
Mean   : 5.513   Mean   : 18.48   Mean   : 30.86   Mean   : 213
3rd Qu.:11.000   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:   0
Max.   :16.000   Max.   :468.00   Max.   :727.00   Max.   :5477
                 NA's   :1        NA's   :1        NA's   :1
```


© 2016 RED PILL Analytics

Look at the NAs at the bottom. Or Hometown has 0's. – Bad Data!

Summarize the Data

```
      TD              Int              Att.1
 Min.   : 0.000   Min.   : 0.0000   Min.   :  0.00
 1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:  0.00
 Median : 0.000   Median : 0.0000   Median :  1.00
 Mean   : 1.269   Mean   : 0.9438   Mean   : 24.95
 3rd Qu.: 0.000   3rd Qu.: 0.0000   3rd Qu.: 12.00
 Max.   :55.000   Max.   :29.0000   Max.   :416.00
 NA's   :1        NA's   :1         NA's   :1
     Yds.1             Y.A              TD.1
 Min.   : -31.0   Min.   :-18.000   Min.   : 0.0000
 1st Qu.:   0.0   1st Qu.:  2.325   1st Qu.: 0.0000
 Median :   0.0   Median :  3.760   Median : 0.0000
 Mean   : 103.3   Mean   :  3.945   Mean   : 0.7353
 3rd Qu.:  47.0   3rd Qu.:  5.000   3rd Qu.: 0.0000
 Max.   :2097.0   Max.   : 60.000   Max.   :28.0000
 NA's   :1        NA's   :3986      NA's   :1
      Rec              Yds.2            Y.R              TD.2
 Min.   :  0.00   Min.   : -15   Min.   :-15.00   Min.   : 0.000
 1st Qu.:  0.00   1st Qu.:   0   1st Qu.:  7.20   1st Qu.: 0.000
 Median :  7.00   Median :  62   Median : 10.04   Median : 0.000
 Mean   : 17.82   Mean   : 206   Mean   : 10.30   Mean   : 1.234
 3rd Qu.: 27.00   3rd Qu.: 285   3rd Qu.: 13.00   3rd Qu.: 2.000
 Max.   :143.00   Max.   :1964   Max.   : 63.00   Max.   :23.000
 NA's   :1        NA's   :1      NA's   :2260     NA's   :1
 FantPos       FantPt        Height..inches.     Weight
 QB:1190   Min.   : -7.00   Min.   :12.00   Min.   :  0.0
 RB:2503   1st Qu.: 10.00   1st Qu.:73.00   1st Qu.:199.0
 TE:1747   Median : 37.00   Median :73.00   Median :217.0
 WR:2758   Mean   : 63.89   Mean   :71.95   Mean   :208.6
           3rd Qu.: 96.75   3rd Qu.:75.00   3rd Qu.:245.0
           Max.   :427.00   Max.   :81.00   Max.   :342.0
           NA's   :1200
```

```
      College              Conference    College.wins
 Miami (FL): 214   Big Ten      :1255   #N/A   :1377
 Michigan  : 197   SEC          :1244   130    : 478
 Florida   : 184   ACC          :1197   112    : 295
 Ohio St.  : 166   Pac 12       :1126   90     : 271
 USC       : 157   Big 12       : 526   86     : 222
 Tennessee : 156   Mountain West: 421   113    : 205
 (Other)   :7124   (Other)      :2429   (Other):5350
 College.losses          DOB          Draft.Round       Draft.Year
 #N/A   :1377   12/30/1899: 276   Min.   : 0.000   Min.   :   0
 74     : 337   3/24/1976 :  30   1st Qu.: 0.000   1st Qu.:   0
 61     : 321   7/18/1900 :  25   Median : 2.000   Median :1999
 86     : 307   2/8/1979  :  23   Mean   : 2.458   Mean   :1363
 58     : 288   6/12/1979 :  21   3rd Qu.: 4.000   3rd Qu.:2004
 80     : 233   6/5/1976  :  21   Max.   :12.000   Max.   :2013
 (Other):5335   (Other)   :7802
   Wonderlic       X40.Yard      Bench.Press    Vert.Leap..in.
 #N/A   :4113   #N/A   :4113   #N/A   :4113   #N/A   :4113
 0      :3402   4.56   : 133   0      :1529   0      : 436
 25     :  61   4.53   : 131   19     : 178   34     : 249
 30     :  55   4.57   : 126   23     : 167   36     : 240
 24     :  51   4.59   : 118   22     : 141   33     : 220
 27     :  46   4.68   : 107   18     : 131   35.5   : 214
 (Other): 470   (Other):3470   (Other):1939   (Other):2726
 Broad.Jump..in.    Shuttle          X3Cone
 #N/A   :4113   #N/A   :4113   #N/A   :4113
 0      : 696   0      :1092   0      :1105
 115    : 247   4.21   : 138   6.96   : 102
 117    : 239   4.16   : 105   7.09   :  87
 118    : 235   4.09   :  93   7      :  81
 114    : 212   4.14   :  84   7.07   :  77
 (Other):2456   (Other):2573   (Other):2633
```

Filter the dataset

- Use Function Nesting to get a subset of data in the summary

```
> summary(subset(nfl, Year==2012))
```

```
      Year                  Player         Age
 Min.   :2012   Alex Smith     :  2   Min.   :21.0
 1st Qu.:2012   Steve Smith    :  2   1st Qu.:24.0
 Median :2012   A.J. Green     :  1   Median :25.0
 Mean   :2012   A.J. Jenkins   :  1   Mean   :26.1
 3rd Qu.:2012   Aaron Hernandez:  1   3rd Qu.:28.0
 Max.   :2012   Aaron Rodgers  :  1   Max.   :38.0
                (Other)        :568
```

```
     Weight
 Min.   :  0.0
 1st Qu.:190.8
 Median :210.0
 Mean   :191.6
 3rd Qu.:240.0
 Max.   :325.0
```



© 2016 RED PILL Analytics

Bad Data?

- If the Mean is 218 for Yards, is it possible to have a max of 5177 or is this bad data?

```
      Yds
Min.   :   0.0
1st Qu.:   0.0
Median :   0.0
Mean   : 218.5
3rd Qu.:   0.0
Max.   :5177.0
```

```
> nfl2012 <- subset(nfl, Year==2012)
> subset(nfl2012, Yds==5177)
   Year     Player Age Hometown Home.State  Tm  G GS Cmp Att  Yds TD
18 2012 Drew Brees  33   Austin         TX NOR 16 16 422 670 5177 43
   Int Att.1 Yds.1  Y.A TD.1 Rec Yds.2 Y.R TD.2 FantPos FantPt
18  19    15     5 0.33    1   0     0  NA    0      QB    346
   Height..inches. Weight College Conference College.wins
18              73      0  Purdue    Big Ten           92
   College.losses        DOB Draft.Round Draft.Year Wonderlic
18             94 12/30/1899           2       2001        28
   X40.Yard Bench.Press Vert.Leap..in. Broad.Jump..in. Shuttle
18     4.83           0             32             105    4.21
   X3Cone
18   7.09
```

Apparently yes. Let's look into at the data by Position.

© 2016 RED PILL Analytics

Group Data by Position

- Here we are grouping with the by function and getting the mean of 4 columns

```
> by(nfl2012[,c("Age","Yds", "Height..inches.", "Weight")], nfl2012$Fant
Pos, colMeans)
nfl2012$FantPos: QB
           Age            Yds Height..inches.         Weight
      27.63889     1745.11111        75.45833        0.00000
------------------------------------------------------------
nfl2012$FantPos: RB
           Age            Yds Height..inches.         Weight
      25.57386        0.00000        71.51705       221.54545
------------------------------------------------------------
nfl2012$FantPos: TE
           Age            Yds Height..inches.         Weight
    26.2436975      0.3277311      76.3445378     250.4201681
------------------------------------------------------------
nfl2012$FantPos: WR
           Age            Yds Height..inches.         Weight
    25.9377990      0.9138756      72.3588517     198.9904306
```



So QBs weigh nothing, that's why they are so agile
Running backs should have yards so that's interesting
TEs are heavy!

© 2016 RED PILL Analytics

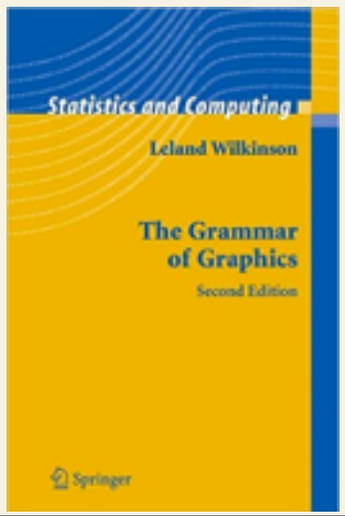


Visualizing Data

Grammar of Graphics Package

- ggplot2 provides many graphing and charting capabilities with R
- Based on Grammar of Graphics by Leland Wilkinson

```
install.packages("ggplot2")
library(ggplot2)
```

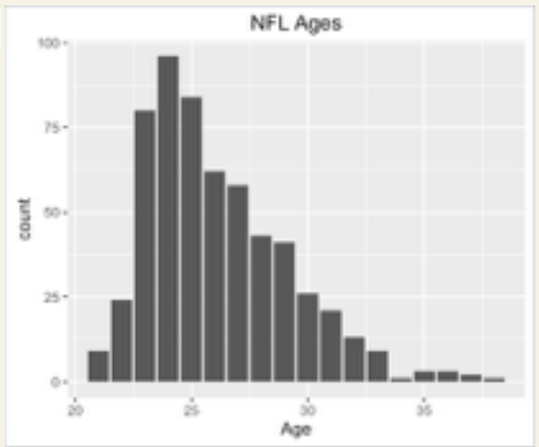


ggplot2 simplifies making complex, multi-layered visualizations

Bar Chart

- Let's view our distribution by Age. Since this is basically discrete data, we'll use a Bar Chart.

```
> qplot(Age, data=nfl2012, geom="bar", main="NFL Ages")
```



Age looks pretty well distributed and accurate. Let's look at other metrics.

© 2016 RED PILL Analytics

Histogram

- Our data imported into R with Factors for some metrics

```
Draft.Year : int 2001 1998 2000 2005 2001 1999 1994 2011 0 0
Wonderlic : Factor w/ 36 levels "#N/A","0","10",..: 2 1 26 1
X40.Yard : Factor w/ 105 levels "#N/A","4.21",..: 26 1 100 5
Bench.Press : Factor w/ 60 levels "#N/A","0","10",..: 11 1 2
Vert.Leap..in. : Factor w/ 43 levels "#N/A","0","24",..: 36
Broad.Jump..in.: Factor w/ 44 levels "#N/A","0","100",..: 2
Shuttle : Factor w/ 102 levels "#N/A","0","3.73",..: 39 1 50
X3Cone : Factor w/ 134 levels "#N/A","0","6.42",..: 32 1 68
```

- Change to Int by converting to a matrix then back to data frame

```
> nfl2012.matrix <- data.matrix(nfl2012)
> nfl2012.mod <- data.frame(nfl2012.matrix)
```

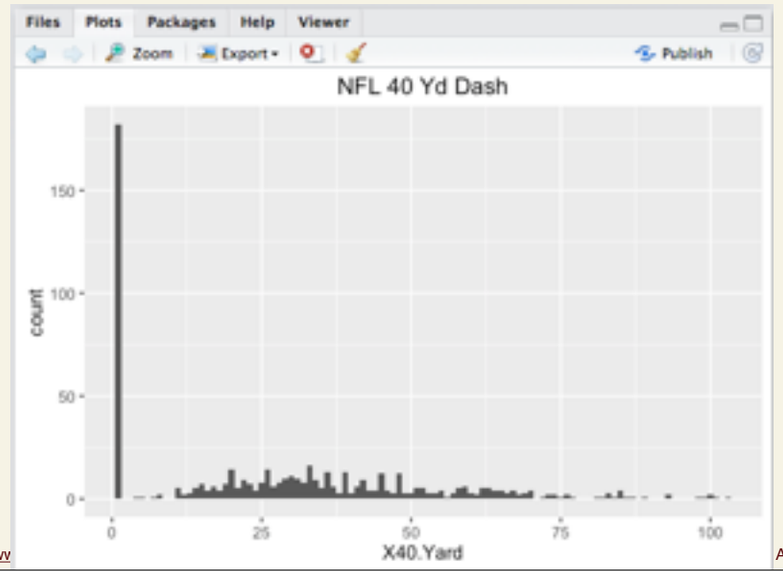
```
Draft.Year : num 2001 2005 2000 2011 2012 ..
Wonderlic : num 21 32 26 14 17 1 9 25 23 21
X40.Yard : num 63 51 100 36 13 1 20 69 81 33
Bench.Press : num 2 2 2 2 2 1 2 2 2 2 ...
Vert.Leap..in. : num 19 24 4 25 33 1 32 2 15
Broad.Jump..in.: num 8 13 44 29 23 1 30 2 8 2
Shuttle : num 39 2 56 36 2 1 58 69 38 27 ...
X3Cone : num 57 87 68 40 2 1 57 88 59 45 ...
```



Factors make R think we have Discrete data not Continuous so we can’t do a histogram on that field

Histogram

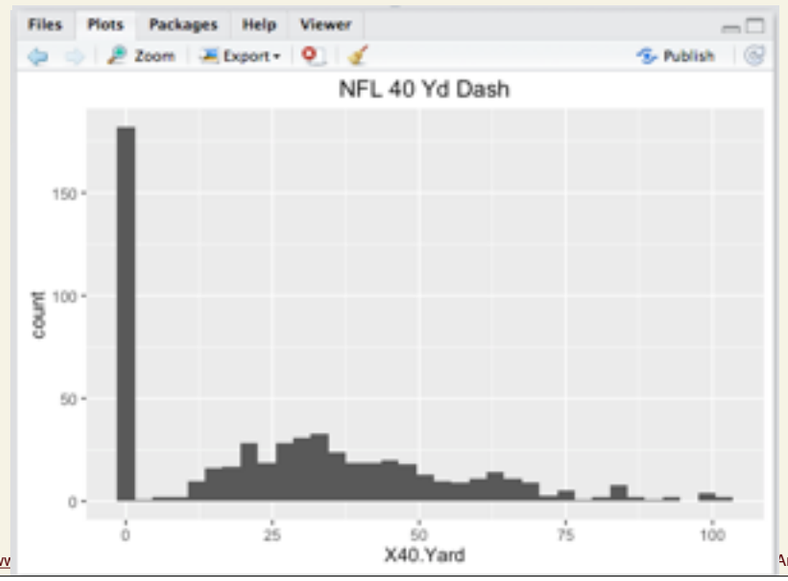
```
> ggplot(nfl2012.mod, aes(x=X40.Yard)) + geom_histogram(binwidth=1)
```



Histograms show us the distribution of the data. We can see that many of the records of this dataset contain low data that is probably not accurate.

Histogram

```
> ggplot(nfl2012.mod, aes(x=X40.Yard)) + geom_histogram(binwidth=3)
```

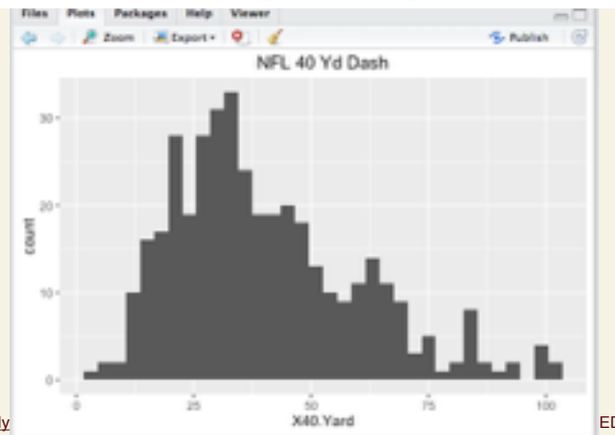


Bins are bigger and easier to see but we have a data problem with so many zeros

Histogram with Some Data Cleanup

- Removed low values

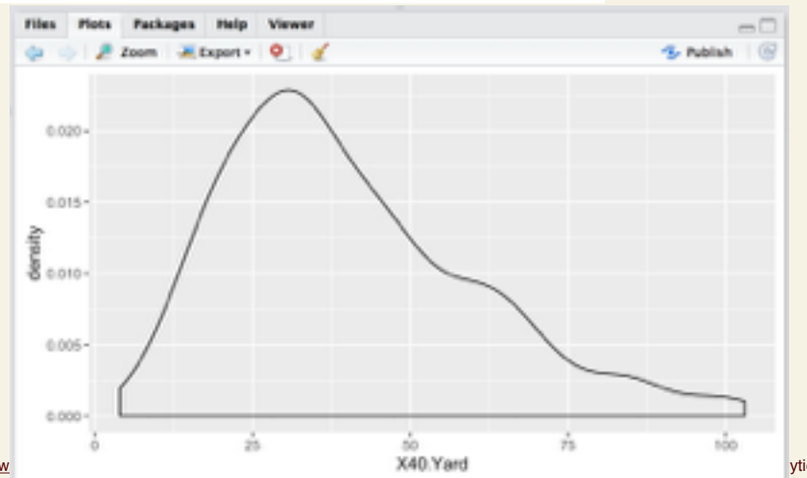
```
> no0in40 <- subset(nfl2012.mod, X40.Yard > 0)
> ggplot(no0in40, aes(x=X40.Yard)) + geom_histogram(binwidth=3)
> no0in40 <- subset(nfl2012.mod, X40.Yard > 1)
> ggplot(no0in40, aes(x=X40.Yard)) + geom_histogram(binwidth=3)
```



Distribution

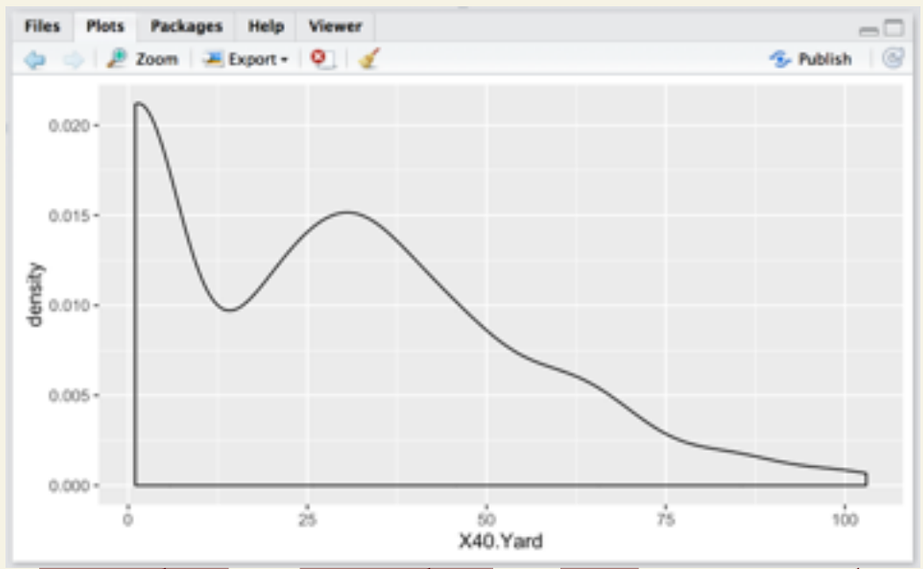
- Density charts are thought to be superior to histograms because you do not need to be concerned with bins

```
> ggplot(no0in40, aes(x=X40.Yard)) + geom_density()
```



Distribution with 0 value data back in

```
> ggplot(nfl2012.mod, aes(x=X40.Yard)) + geom_density()
```



Quick Clean Up

rm removes a variable or dataset

```
> rm(no0in40)
```

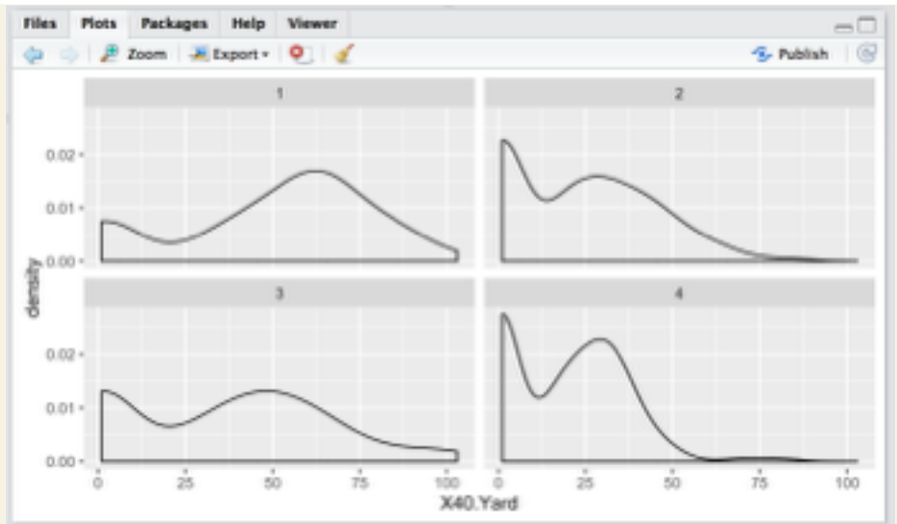


 © 2016 RED PILL Analytics

Group the Chart by a Dimension

- We can add a “facet wrap” to group our charts by a dimension

```
> ggplot(nfl2012.mod, aes(x=X40.Yard)) + geom_density() + facet_wrap(~ FantPos)
```



Note that when we changed to a matrix, it changed the text values of Position to numbers. There’s workarounds to this but for the simplicity of this demonstration, know that 1 = QB, 2 = RB, 3 = TE, 4 = WR

Distributions for Categorical Data

- Can get a count of how many records exist for each value in a table format

```
> xtabs(~ FantPos, data=nfl2012)
FantPos
 QB  RB  TE  WR
 72 176 119 209
> xtabs(~ Draft.Year, data=nfl2012)
Draft.Year
   0 1997 1998 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
 200    1    4    2    7    5   10   17   20   30   28   41   44   42   64   61
```



© 2016 RED PILL Analytics

Distribution for 2 data points

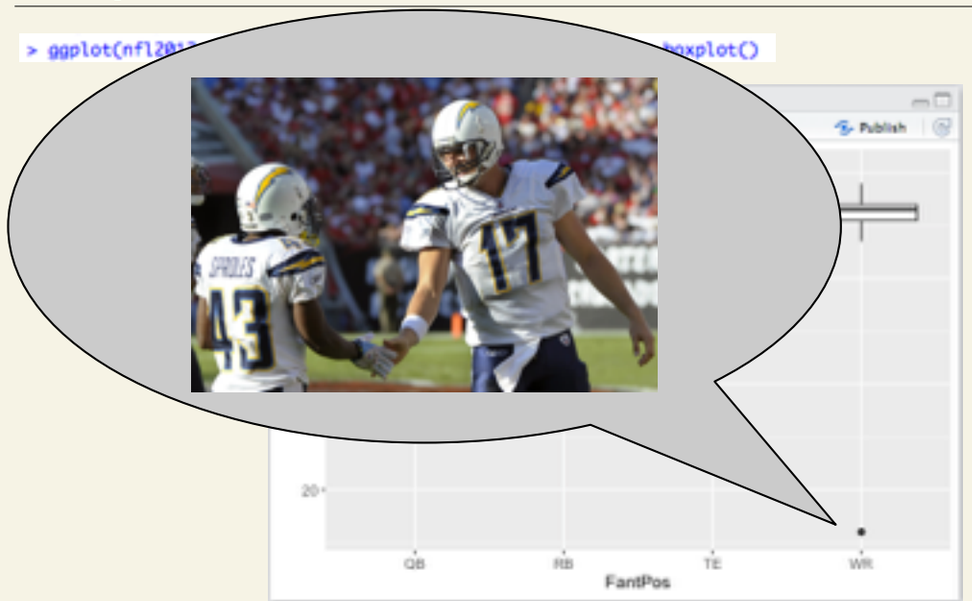
- Can change this to a 2 way cross tab distribution

```
> xtabs(~ FantPos + Draft.Year, data=nfl2012)
       Draft.Year
FantPos  0 1997 1998 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
     QB 13    0    3    1    2    1    2    5    8    5    3    4    3    4   10    8
     RB 68    0    0    0    0    0    1    4    5    7    9   14   15   11   24   18
     TE 38    1    0    0    1    0    2    4    4    8    7   11    9   12   10   12
     WR 81    0    1    1    4    4    5    4    3   10    9   12   17   15   20   23
```



 © 2016 RED PILL Analytics

Boxplot



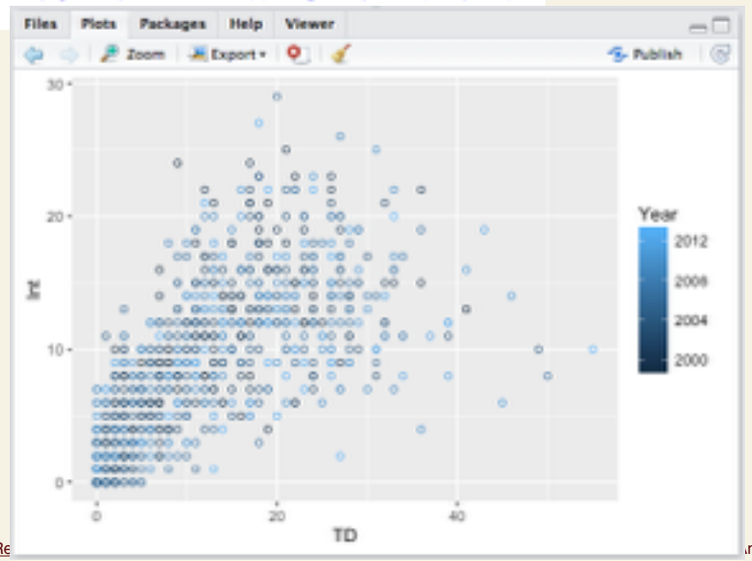
Boxplot shows us distribution by quartiles. It also helps identify outliers.

What's going on with that WR who's so short??

 © 2016 RED PILL Analytics

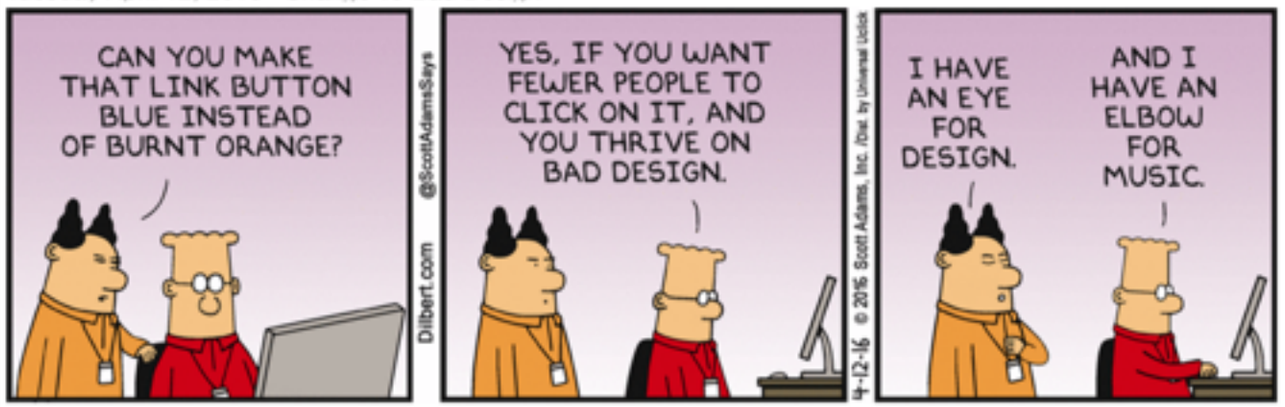
Scatterplot

```
> qbs <- subset(nfl, FantPos=="QB")
> ggplot(qbs, aes(x=TD, y=Int, color=Year)) + geom_point(shape=1)
> rm(qbs)
```



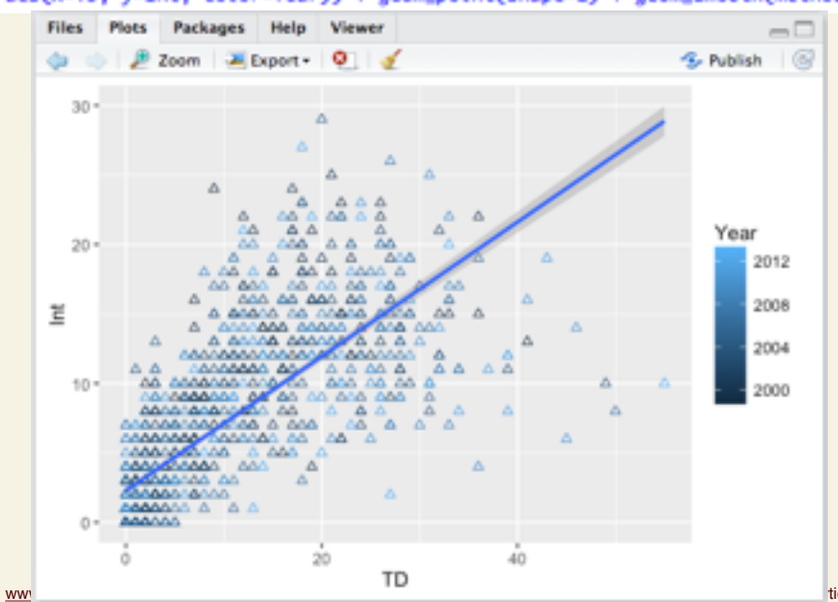
Notice the addition of color. This can be added to any chart you've already seen.

Tuesday April 12, 2016 *Change To Bad Design*



Scatterplot with Regression

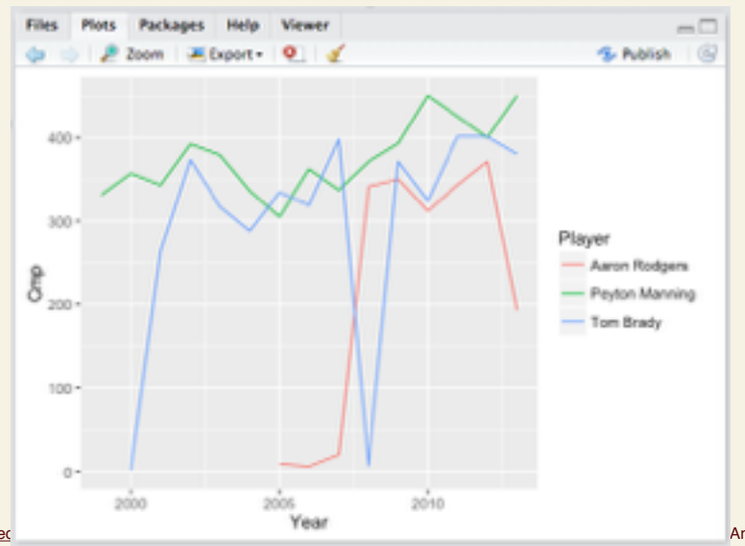
```
> ggplot(qbs, aes(x=TD, y=Int, color=Year)) + geom_point(shape=2) + geom_smooth(method=lm)
```



Defaults to 95% confidence
Shaded area is confidence region

Line Chart

```
> myQBs <- subset(nfl, Player %in% c("Aaron Rodgers", "Peyton Manning", "Tom Brady"))
> ggplot(data=myQBs, aes(x=Year, y=Cmp, group=Player, color=Player)) + geom_line()
```



First there's too many players in my dataset to make a good line chart. Filter down to 3 quarterbacks. Which one is the best? Well the only correct answer is Rodgers. Doesn't matter what data says.

Add a Bar Chart to the Line

```
> ggplot(data=myQBs, aes(x=Year, y=TD, fill=Player)) + geom_bar(stat="identity") + geom_line
(data=myQBs, aes(x=Year, y=Cmp, color=Player, group=Player)) + ylab("TD & Completions")
```



Let’s step through each of the pieces here

Stacked Bars are Rarely Helpful

```
> ggplot(data=myQBs, aes(x=Year, y=TD, fill=Player)) + geom_bar(stat="identity", position="d
odge") + geom_line(data=myQBs, aes(x=Year, y=Cmp, color=Player, group=Player)) + ylab("TD &
Completions")
```



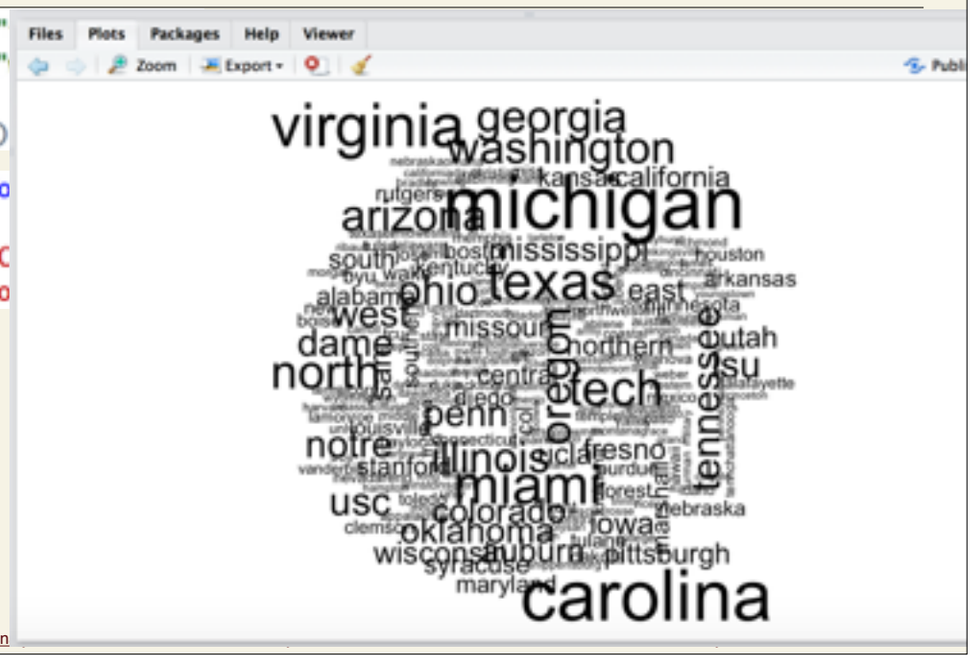
Can’t compare between the values easily
Added position=“dodge”


© 2016 RED PILL Analytics
What about Text fields?

Word cloud

```
install.packages("
install.packages("
library(tm)
library(wordcloud)
```

```
> wordcloud(nfl$Co
Warning message:
In wordcloud(nfl$C
  florida could no
```



Very easy to do in R!


© 2016 RED PILL Analytics
Missing Data

Null vs NA in R

R treats NA like other languages consider NULL

	NULL	NA
Definition	Null object, a reserved word	Logical constant of length 1 containing a missing value indicator
Behavior in Vector	Not allowed. Won't save within vector.	Exists and represents missing value.
Behavior in List (such as Data Frame)	Can exist if not assigned but created with it.	Exists and represents missing value.



© 2016 RED PILL Analytics

Nulls on Import

Our dataset had nulls in it when we pulled it into R. How were they assigned?

```
> subset(nfl, Player=="Jeff Garcia")
     Year      Player Age Hometown Home.State  Tm  G GS Cmp Att
24   2000 Jeff Garcia  30   Gilroy         CA SFO 16 16 355 561
55   2001 Jeff Garcia  31   Gilroy         CA SFO 16 16 316 504
151  2002 Jeff Garcia  32   Gilroy         CA SFO 16 16 328 528
246  2003 Jeff Garcia  33   Gilroy         CA SFO 13 13 225 392
738  2008 Jeff Garcia  38   Gilroy         CA TAM 12 11 244 376
815  1999 Jeff Garcia  29   Gilroy         CA SFO 13 10 225 375
847  2007 Jeff Garcia  37   Gilroy         CA TAM 13 13 209 327
1318 2004 Jeff Garcia  34   Gilroy         CA CLE 11 10 144 252
1700 2006 Jeff Garcia  36   Gilroy         CA PHI  8  6 116 188
3028 2005 Jeff Garcia  35   Gilroy         CA DET  6  5 102 173
7871 2009 Jeff Garcia  39   Gilroy         CA PHI  1  0   0   0
      Yds TD Int Att.1 Yds.1   Y.A TD.1 Rec Yds.2 Y.R TD.1 PlantPos
24   4278 31  10    72   414  5.75    4   0     0  NA
55   3538 32  12    72   254  3.53    5   0     0  NA           QB
151  3344 21  10    73   353  4.84    3   0     0  NA    0      QB
246  2704 18  13    56   319  5.70    7   1     6   6    0      QB
738  2712 12   6    35   148  4.23    1   0     0  NA    0      QB
815  2544 11  11    45   231  5.13    2   0     0  NA    0      QB
847  2440 13   4    35   116  3.31    1   0     0  NA    0      QB
1318 1731 10   9    35   169  4.83    2   0     0  NA    0      QB
```

© 2016 RED PILL Analytics

Finding Missing Data

```
> na_count <-sapply(nfl, function(y) sum(is.na(y)))
> na_count$name<-rownames(na_count)
Warning message:
In na_count$name <- rownames(na_count) : Coercing LHS to a list
> View(na_count)
```



	Year	Player	Age	Hometown	Home.State	Tm	G	GS	Cmp	Att	Yds	TD	Int	Att.1	Yds.1	Y.A	TD.1	Rec	Yds.2	Y.R	TD.2	FantPos	Fa
1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	3986	1	1	1	2260	1	0	12

This will count the number of NA records in each column.
The second line assigns the column names to each column in our new data frame.

© 2016 RED PILL Analytics

But look what else we found in Jeff's records!

Console ~/

```
24          72          107 2/24/1970          0          0
55          72          107 2/24/1970          0          0
151         72          107 2/24/1970          0          0
246         72          107 2/24/1970          0          0
738         72          107 2/24/1970          0          0
815         72          107 2/24/1970          0          0
847         72          107 2/24/1970          0          0
1318        72          107 2/24/1970          0          0
1700        72          107 2/24/1970          0          0
3028        72          107 2/24/1970          0          0
7871        72          107 2/24/1970          0          0
     Wonderlic X40.Yard Bench.Press Vert.Leap..in. Broad.Jump..in.
24        #N/A     #N/A        #N/A                          #N/A
55        #N/A     #N/A        #N/A                          #N/A
151       #N/A     #N/A        #N/A                          #N/A
246       #N/A     #N/A        #N/A                          #N/A
738       #N/A     #N/A        #N/A                          #N/A
815       #N/A     #N/A        #N/A                          #N/A
847       #N/A     #N/A        #N/A                          #N/A
1318      #N/A     #N/A        #N/A                          #N/A
1700      #N/A     #N/A        #N/A                          #N/A
3028      #N/A     #N/A        #N/A          #N/A            #N/A
7871      #N/A     #N/A        #N/A          #N/A            #N/A
     Shuttle X3Cone
```



The data is this way in the file. We should make all missing values consistent.

Make Missing Data Consistent in R

```
> nfl[nfl=="#N/A"] = NA
> sapply(subset(nfl, Player=="Jeff Garcia"), function(y) sum(is.na(y)))
            Year           Player              Age         Hometown       Home.State
               0                0                0                0                0
              Tm                G               GS              Cmp              Att
               0                0                0                0                0
             Yds               TD              Int           Att.1            Yds.1
               0                0                0                0                0
             Y.A             TD.1              Rec            Yds.2              Y.R
               0                0                0                0               10
            TD.2          FantPos           FantPt Height..inches.           Weight
               0                0                1                0                0
         College       Conference     College.wins  College.losses              DOB
               0                0                0                0                0
    Draft.Round       Draft.Year        Wonderlic         X40.Yard       Bench.Press
               0                0               11               11               11
 Vert.Leap..in. Broad.Jump..in.          Shuttle           X3Cone
              11               11               11               11
```


Second line is same function as the nacheck previously but displaying in console instead of storing results

Check the whole dataset now

```
> sapply(nfl, function(y) sum(is.na(y)))
            Year           Player              Age         Hometown       Home.State
               0                0                0                0                0
              Tm                G               GS              Cmp              Att
               0                0                0                1                1
             Yds               TD              Int           Att.1            Yds.1
               1                1                1                1                1
             Y.A             TD.1              Rec           Yds.2              Y.R
            3986                1                1                1             2260
            TD.2          FantPos           FantPt Height..inches.           Weight
               1                0             1200                0                0
         College       Conference     College.wins   College.losses              DOB
              20              375             1377             1377                0
    Draft.Round      Draft.Year        Wonderlic         X40.Yard      Bench.Press
               0                0             4113             4113             4113
 Vert.Leap..in. Broad.Jump..in.          Shuttle           X3Cone
            4113             4113             4113             4113
```

 © 2016 RED PILL Analytics

© 2016 RED PILL Analytics



What to do about missing & bad data?

Depends on dataset, company, requirements. Needs to be defined for each case.

Handling Bad Data in ETL



In some instances, the data might not exist anywhere.
In others, you might be able to clean it up (mis-spellings, use lookup table, etc).
Decide if these rows should be excluded from analysis since they are missing data. Should ETL reject them? Should anyone be notified when bad data comes thru ETL jobs?

© 2016 RED PILL Analytics



Using Data Quality Package

DataQualityR Package

```
> install.packages("dataQualityR")
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0  0 24
859     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0100 24859  1
00 24859    0     0  27705      0 --:--:-- --:--:-- --:--:-- 27744

The downloaded binary packages are in

/var/folders/5d/wmywfynn5ss98dn0zy_524080000gn/T//RtmpJJFUKz/downloaded_packages
> library("dataQualityR")
```

```
> checkDataQuality(data=nfl, out.file.num= "/users/michellekolbe/Desktop/dq_nfl_num.csv", out.file.cat= "/users/michellekolbe/Desktop/dq_nfl_cat.csv")
Check for numeric variables completed // Results saved to disk // Time difference of 0.08018517 secs
Check for categorical variables completed // Results saved to disk // Time difference of 0.126879 secs
```


© 2016 RED PILL Analytics

Numerical Results

Table 1

	non-missing	missing	missing percent	unique	mean	min	p1	p5	p10	p25	p50	p75	p90	p95	p99	max
Year	8198	0	0	15	2006.13	1999	1999	1999	2000	2002	2006	2010	2012	2013	2013	2013
Age	8198	0	0	24	26.65	21	22	22	23	24	26	29	32	33	37	44
G	8198	0	0	17	11.23	0	1	1	2	7	14	16	16	16	16	16
GS	8198	0	0	17	5.51	0	0	0	0	0	3	11	15	16	16	16
Cmp	8197	1	0.01	350	18.48	0	0	0	0	0	0	0	18	171	341	468
Att	8197	1	0.01	480	30.86	0	0	0	0	0	0	0	32	292	548.04	727
Yds	8197	1	0.01	900	213.02	-6	0	0	0	0	0	0	201	1921.6	4008.04	5477
TD	8197	1	0.01	46	1.27	0	0	0	0	0	0	0	1	10.2	27	55
Int	8197	1	0.01	30	0.94	0	0	0	0	0	0	0	1	9	17	29
Att.1	8197	1	0.01	346	24.95	0	0	0	0	0	1	12	74	175	314	416
Yds.1	8197	1	0.01	906	103.29	-31	-6	0	0	0	0	47	316.4	719.2	1339.04	2097
Y.A	4212	3986	48.62	769	3.94	-18	-6	-1	0.13	2.33	3.76	5	7.5	10	17	60
TD.1	8197	1	0.01	24	0.74	0	0	0	0	0	0	0	2	5	12	28
Rec	8197	1	0.01	121	17.82	0	0	0	0	0	7	27	53	79	94.04	143
Yds.2	8197	1	0.01	1135	206	-15	0	0	0	0	62	285	658.4	905.2	1325	1964
Y.R	5938	2260	27.57	1391	10.3	-15	0	4	5.1	7.2	10.04	13	15.62	17.45	23	63
TD.2	8197	1	0.01	20	1.23	0	0	0	0	0	0	2	4	6	10	23
FantPt	6998	1200	14.64	338	63.89	-7	1	1	3	10	37	96.75	170	214	290	427
Height..inches.	8198	0	0	17	71.95	12	12	68	70	73	73	75	77	77	79	81
Weight	8198	0	0	140	208.57	0	0	0	182	199	217	245	257	263	278	342
Draft.Round	8198	0	0	13	2.46	0	0	0	0	0	2	4	6	7	7	12
Draft.Year	8198	0	0	32	1362.53	0	0	0	0	0	1999	2004	2008	2010	2012	2013

Categorical Results

	n.non.miss	n.miss	n.miss.percent	n.unique	cat_1	freq_1	cat_2	freq_2	cat_3	freq_3	cat_4	freq_
Player	8198	0	0	2302	Steve Smith	18	Alex Smith	17	Adrian Peterson	15	Brandon Stokley	1
Hometown	8198	0	0	952	0	226	Houston	143	Miami	143	New Orleans	13
Home.State	7972	226	2.76	67	CA	1081	FL	746	TX	717	LA	41
Tm	8198	0	0	36	TAM	279	GNB	272	ARI	268	DEN	26
FantPos	8198	0	0	4	WR	2758	RB	2503	TE	1747	QB	119
College	8178	20	0.24	353	Miami (FL)	214	Michigan	197	Florida	184	Ohio St.	16
Conference	7822	376	4.59	43	Big Ten	1255	SEC	1244	ACC	1197	Pac 12	112
College.wins	6821	1377	16.8	76	130	478	112	295	90	271	86	22
College.losses	6821	1377	16.8	67	74	337	61	321	86	307	58	28
DOB	8198	0	0	1970	12/30/1899	276	3/24/1976	30	7/18/1900	25	2/8/1979	2
Wonderlic	4085	4113	50.17	36	0	3402	25	61	30	55	24	5
X40.Yard	4085	4113	50.17	105	4.56	133	4.53	131	4.57	126	4.59	11
Bench.Press	4085	4113	50.17	60	0	1529	19	178	23	167	22	14
Vert.Leap..in.	4085	4113	50.17	43	0	436	34	249	36	240	33	22
Broad.Jump..in.	4085	4113	50.17	44	0	696	115	247	117	239	118	23
Shuttle	4085	4113	50.17	102	0	1092	4.21	138	4.16	105	4.09	9
X3Cone	4085	4113	50.17	134	0	1105	6.96	102	7.09	87	7	8

In Summary

- R gives you a quick and easy way to learn about your data before investing time into ETL
- Open source means no investment into tools
- R isn't scary or all statistical and stuff!

© 2016 RED PILL Analytics


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY ARE THERE SQUIRRELS
WHY DO YOUR BOOBS HURT
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY IS THERE LAVA
WHY ARE THERE FEMAL
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PARADES ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE MY BOOBS ITCHY
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS THERE NO KING IN
WHY ARE DOGS AFRAID OF


come join in our community
IRISH PUB
RESTAURANT
BAR
MUSIC ROOM
ODTUG EPM - BI - Database Community Meetup
WHERE:
Ri Ra Irish Pub
Located in Mandalay Bay, The Parlour Room
The Shoppes at Mandalay Bay Place
3930 Las Vegas Blvd South
Las Vegas, NV 89119
WHEN:
Tuesday, April 12, 7:00 PM
ODTUG

Using R for Data Profiling
Session #1805

Fill out a session survey in the mobile app!!

Michelle Kolbe

medium.com/@datacheesehead

@mekolbe

linkedin.com/in/michellekolbe

michelle.kolbe@redpillanalytics.com

© 2016 RED PILL Analytics