MANGOSOLUTIONS
data analysis that delivers

Analytics Projects in R in a Nutshell: Everything You Need to Know

Nic Crane and Sam Cartwright

Overview

- Stages of an analytics project one-by-one
  - What? Why? How?

- Example dataset
  - Titanic – widely used in ML tutorials

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Titanic Dataset

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500	NA	S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250	NA	S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500	NA	S
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583	NA	Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750	NA	S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333	NA	S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708	NA	C
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7000	G6	S
12	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.5500	C103	S
13	13	0	3	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151	8.0500	NA	S
14	14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	31.2750	NA	S
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	7.8542	NA	S
16	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	16.0000	NA	S
17	17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.1250	NA	Q
18	18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0000	NA	S

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Kinds of Analytics Problems

Descriptive Analytics
What happened?

Diagnostic Analytics
Why did that happen?

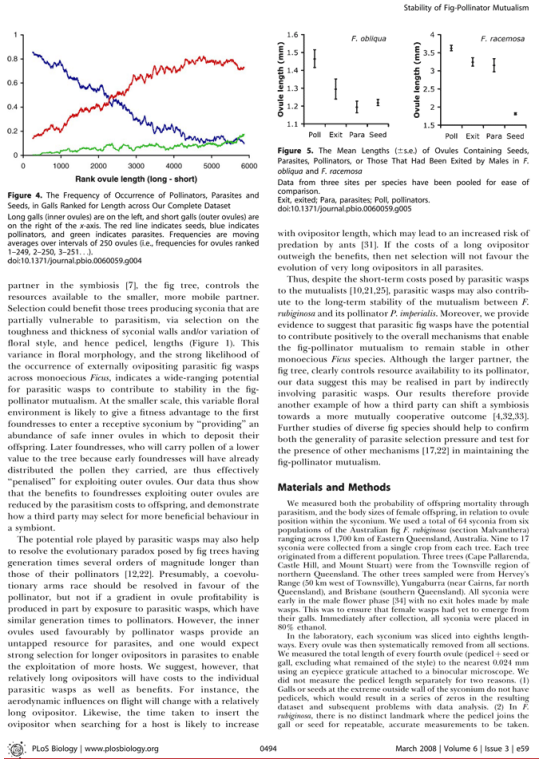
Predictive Analytics
What will happen?

Prescriptive Analytics
“Best” course of action?

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Descriptive and Diagnostic Analytics

- What happened?
- Why did it happen?
- e.g. reporting the results of an experiment



Stability of Fig-Pollinator Mutualism

Figure 4. The Frequency of Occurrence of Pollinators, Parasites and Seeds, in Galls Ranked for Length across Our Complete Dataset
Long galls (inner ovules) are on the left, and short galls (outer ovules) are on the right of the *x*-axis. The red line indicates seeds, blue indicates pollinators, and green indicates parasites. Frequencies are moving averages over intervals of 250 ovules (i.e., frequencies for ovules ranked 1–249, 2–250, 3–251 …).
doi:10.1371/journal.pbio.0060059.g004

Figure 5. The Mean Lengths (±s.e.) of Ovules Containing Seeds, Parasites, Pollinators, or Those That Had Been Exited by Males in *F. obliqua* and *F. racemosa*
Data from three sites per species have been pooled for ease of comparison.
Exit, exited; Para, parasites; Poll, pollinators.
doi:10.1371/journal.pbio.0060059.g005

partner in the symbiosis [7], the fig tree, controls the resources available to the smaller, more mobile partner. Selection could benefit those trees producing syconia that are partially vulnerable to parasitism, via selection on the toughness and thickness of syconial walls and/or variation of floral style, and hence pedicel, lengths (Figure 1). This variance in floral morphology, and the strong likelihood of the occurrence of externally ovipositing parasitic fig wasps across monoecious *Ficus*, indicates a wide-ranging potential for parasitic wasps to contribute to stability in the fig-pollinator mutualism. At the smaller scale, this variable floral environment is likely to give a fitness advantage to the first foundresses to enter a receptive syconium by "providing" an abundance of safe inner ovules in which to deposit their offspring. Later foundresses, who will carry pollen of a lower value to the tree because early foundresses will have already distributed the pollen they carried, are thus effectively "penalised" for exploiting outer ovules. Our data thus show that the benefits to foundresses exploiting outer ovules are reduced by the parasitism costs to offspring, and demonstrate how a third party may select for more beneficial behaviour in a symbiont.

The potential role played by parasitic wasps may also help to resolve the evolutionary paradox posed by fig trees having generation times several orders of magnitude longer than those of their pollinators [12,22]. Presumably, a coevolutionary arms race should be resolved in favour of the pollinator, but not if a gradient in ovule profitability is produced in part by exposure to parasitic wasps, which have similar generation times to pollinators. However, the inner ovules used favourably by pollinator wasps provide an untapped resource for parasites, and one would expect strong selection for longer ovipositors in parasites to enable the exploitation of more hosts. We suggest, however, that relatively long ovipositors will have costs to the individual parasitic wasps as well as benefits. For instance, the aerodynamic influences on flight will change with a relatively long ovipositor. Likewise, the time taken to insert the ovipositor when searching for a host is likely to increase with ovipositor length, which may lead to an increased risk of predation by ants [31]. If the costs of a long ovipositor outweigh the benefits, then net selection will not favour the evolution of very long ovipositors in all parasites.

Thus, despite the short-term costs posed by parasitic wasps to the mutualists [10,21,25], parasitic wasps may also contribute to the long-term stability of the mutualism between *F. rubiginosa* and its pollinator *P. imperialis*. Moreover, we provide evidence to suggest that parasitic fig wasps have the potential to contribute positively to the overall mechanisms that enable the fig-pollinator mutualism to remain stable in other monoecious *Ficus* species. Although the larger partner, the fig tree, clearly controls resource availability to its pollinator, our data suggest this may be realised in part by indirectly involving parasitic wasps. Our results therefore provide another example of how a third party can shift a symbiosis towards a more mutually cooperative outcome [4,32,33]. Further studies of diverse fig species should help to confirm both the generality of parasite selection pressure and test for the presence of other mechanisms [17,22] in maintaining the fig-pollinator mutualism.

Materials and Methods

We measured both the probability of offspring mortality through parasitism, and the body sizes of female offspring, in relation to ovule position within the syconium. We used a total of 64 syconia from six populations of the Australian fig *F. rubiginosa* (section Malvanthera) ranging across 1,700 km of Eastern Queensland, Australia. Nine to 17 syconia were collected from a single crop from each tree. Each tree originated from a different population. Three trees (Cape Palarenda, Castle Hill, and Mount Stuart) were from the Townsville region of northern Queensland. The other trees sampled were from Hervey's Range (50 km west of Townsville), Yungaburra (near Cairns, far north Queensland), and Brisbane (southern Queensland). All syconia were early in the male flower phase [34] with no exit holes made by male wasps. This was to ensure that female wasps had yet to emerge from their galls. Immediately after collection, all syconia were placed in 80% ethanol.

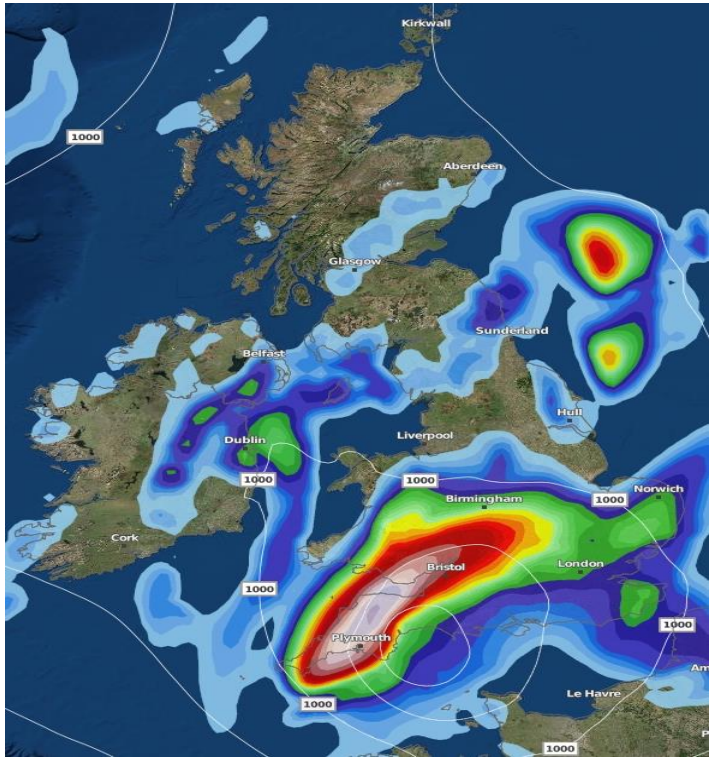
In the laboratory, each syconium was sliced into eighths lengthways. Every ovule was then systematically removed from all sections. We measured the total length of every fourth ovule (pedicel + seed or gall, excluding what remained of the style) to the nearest 0.024 mm using an eyepiece graticule attached to a binocular microscope. We did not measure the pedicel length separately for two reasons. (1) Galls or seeds at the extreme outside wall of the syconium do not have pedicels, which would result in a series of zeros in the resulting dataset and subsequent problems with data analysis. (2) In *F. rubiginosa*, there is no distinct landmark where the pedicel joins the gall or seed for repeatable, accurate measurements to be taken.

MANGOSOLUTIONS
data analysis that delivers

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Predictive Analytics

- What will happen?
- What action should we take?
- e.g. weather, parking spaces in Bath



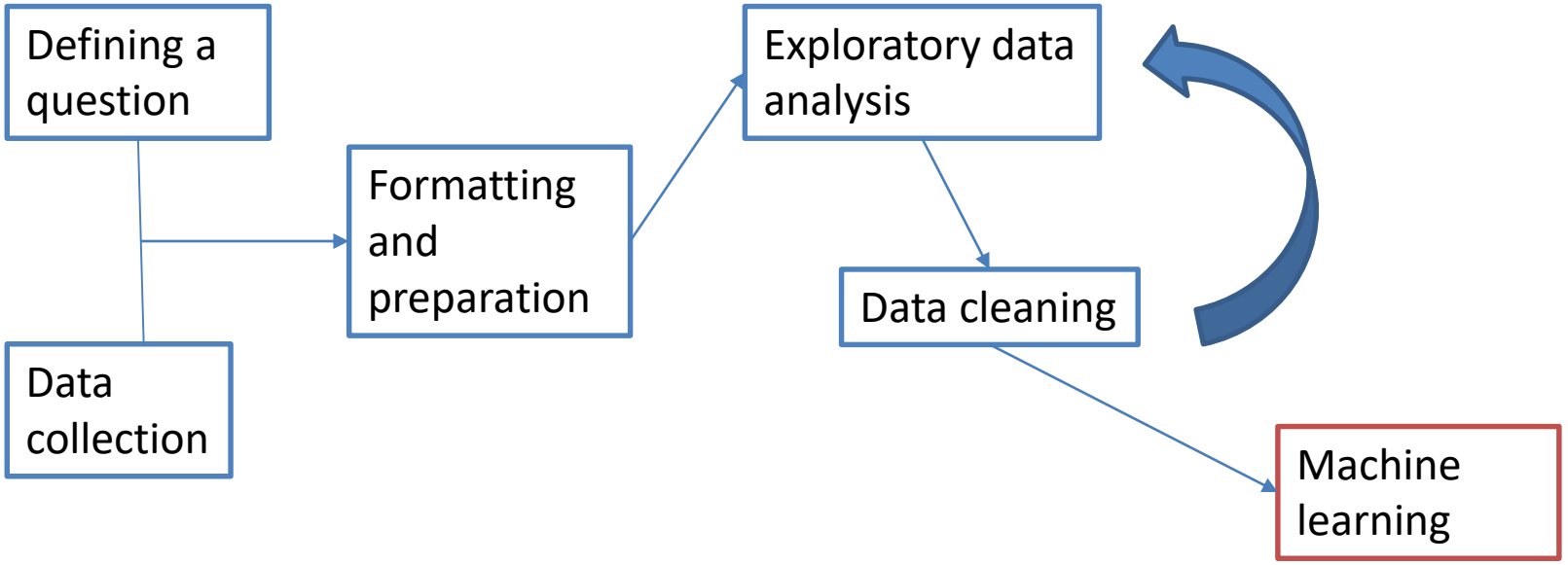
Nic Crane – Data Scientist
ncrane@mango-solutions.com

What kind of question?

- Regression
- Classification
- Something else…

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Stages of analysis



In reality there is no strict order of stages and you will go back and forth as your knowledge increases.

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Defining a question

- Datasets are often rich with information
- Rich enough to answer a huge amount of questions
- Defining a question will help streamline your analysis
- Each step should lead you closer to the question

Nic Crane – Data Scientist
ncrane@mango-solutions.com

COLLECTING DATA

MANGOSOLUTIONS
data analysis that delivers

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Downloading

- Manually download

- Download directly into R
  - download.file(url_to_file)
  - unzip(zip_file)

- **readr** - import csv, fwf, tsv, and other formats
- **readxl** - import Excel files
- **haven** - import SPSS, Stata and SAS files.

Nic Crane – Data Scientist
ncrane@mango-solutions.com

readr example

```
> library(readr)
> train <- read_csv("http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/train.csv")
Parsed with column specification:
cols(
  PassengerId = col_integer(),
  Survived = col_integer(),
  Pclass = col_integer(),
  Name = col_character(),
  Sex = col_character(),
  Age = col_double(),
  SibSp = col_integer(),
  Parch = col_integer(),
  Ticket = col_character(),
  Fare = col_double(),
  Cabin = col_character(),
  Embarked = col_character()
)
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com

Databases in R

- **DBI** – allows to connect to databases
- Individual packages for functionality related to that database, e.g. **RSQLite**, **RMySQL** etc

Nic Crane – Data Scientist
ncrane@mango-solutions.com

APIs

- Let applications communicate
- Exposes internal functions
- Google, Facebook, Twitter, Dropbox…loads of APIs available out there

Nic Crane – Data Scientist
ncrane@mango-solutions.com

API Example

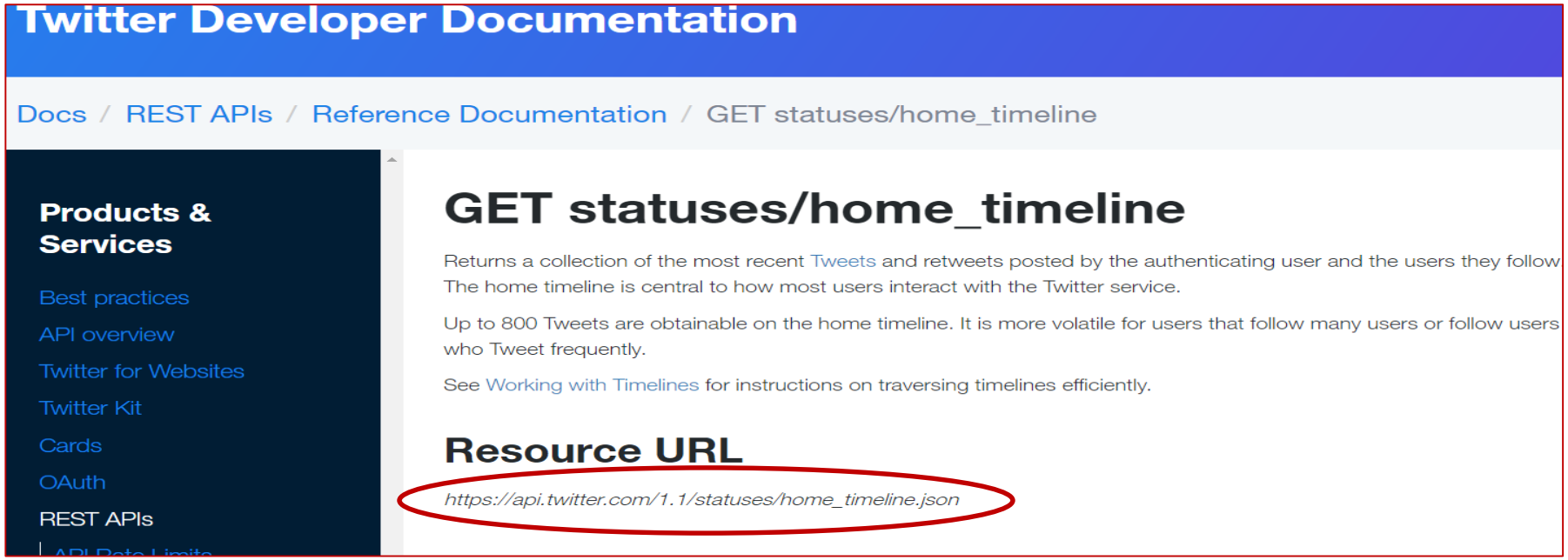
http://thisisnic.github.io/Twitter-in-R/twitter_API.html

Steps:

1. Sign up to use API
2. Get credentials for using API
3. Access API using oauth
4. Retrieve content

Nic Crane – Data Scientist
ncrane@mango-solutions.com

API Example




```
my_timeline=GET("https://api.twitter.com/1.1/statuses/home_timeline.json", sig)
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com

APIs in R

- **httr** for authentication and getting data
- **jsonlite** for parsing data from JSON format into R data object

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Web Scraping

- Extracting data directly from website
- Useful when no API available
- Used for price comparison websites

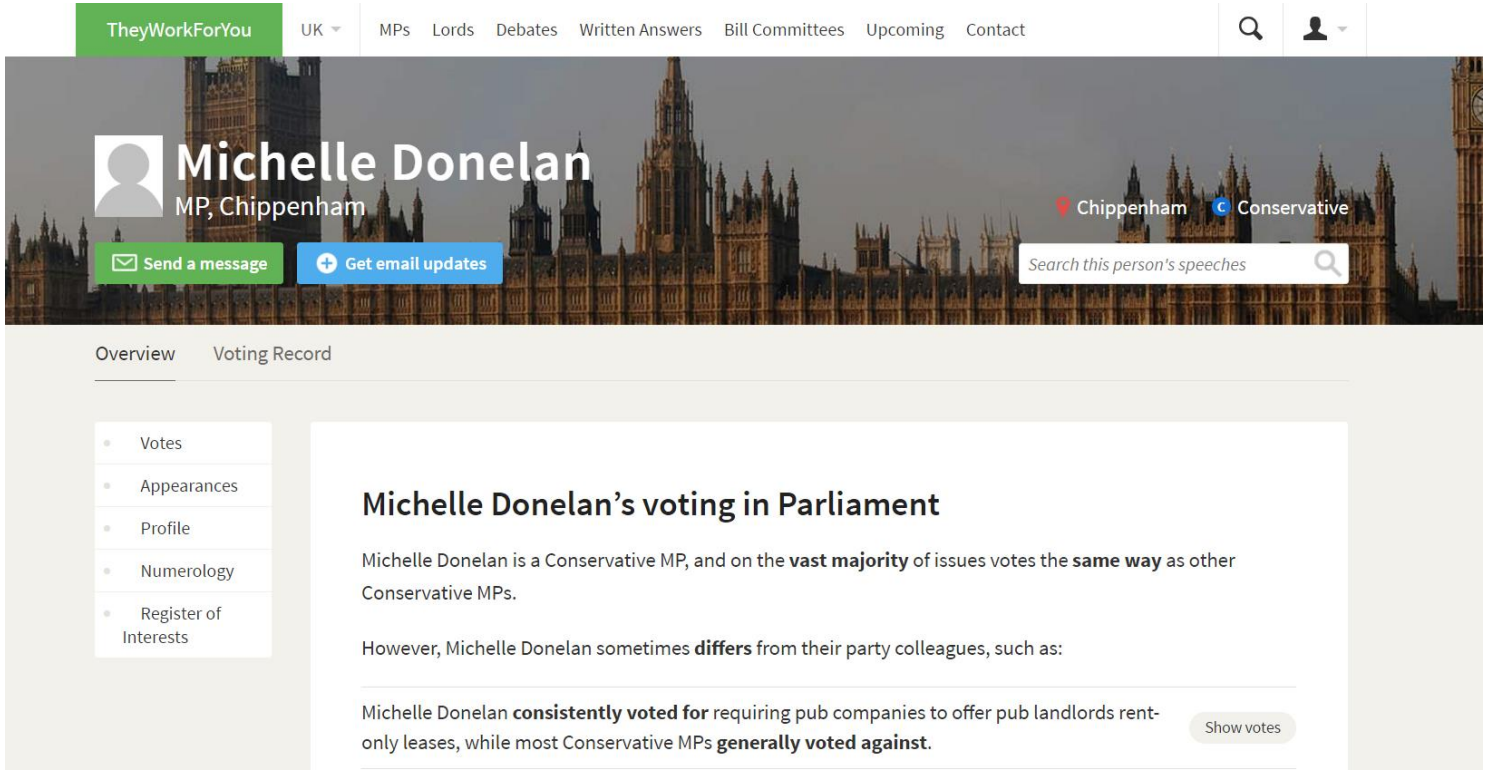
Nic Crane – Data Scientist
ncrane@mango-solutions.com

Web Scraping in R

- **rvest**
- SelectorGadget - bookmarklet

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Web Scraping



Nic Crane – Data Scientist
ncrane@mango-solutions.com

Data Formatting

- Tidy data
- **tidyr**
- **dplyr**

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Tidy Data

The two most important properties of tidy data are:

- Each column is a variable.
- Each row is an observation.

See: http://vita.had.co.nz/papers/tidy-data.html

Nic Crane – Data Scientist
ncrane@mango-solutions.com

tidyr

Key functionality:

- Wide format to long format and vice versa
- Multiple columns to single columns and vice versa

Nic Crane – Data Scientist
ncrane@mango-solutions.com

dplyr

- Data manipulation
- Use functions instead of square brackets and dollar signs
- Compatible with databases

Nic Crane – Data Scientist
ncrane@mango-solutions.com

dplyr Example

```
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

> filter(train, Age < 10)
# A tibble: 62 x 12
   PassengerId Survived Pclass                                      Name    Sex   Age SibSp Parch        Ticket    Fare Cabin Embarked
         <int>    <int>  <int>                                     <chr>  <chr> <dbl> <int> <int>         <chr>   <dbl> <chr>    <chr>
1            8        0      3          Palsson, Master. Gosta Leonard   male  2.00     3     1        349909 21.0750  <NA>        S
2           11        1      3         Sandstrom, Miss. Marguerite Rut female  4.00     1     1       PP 9549 16.7000    G6        S
3           17        0      3                     Rice, Master. Eugene   male  2.00     4     1        382652 29.1250  <NA>        Q
4           25        0      3           Palsson, Miss. Torborg Danira female  8.00     3     1        349909 21.0750  <NA>        S
5           44        1      2 Laroche, Miss. Simonne Marie Anne Andree female  3.00     1     2 SC/Paris 2123 41.5792  <NA>        C
6           51        0      3               Panula, Master. Juha Niilo   male  7.00     4     1       3101295 39.6875  <NA>        S
7           59        1      2             West, Miss. Constance Mirium female  5.00     1     2    C.A. 34651 27.7500  <NA>        S
8           64        0      3                    Skoog, Master. Harald   male  4.00     3     2        347088 27.9000  <NA>        S
9           79        1      2           Caldwell, Master. Alden Gates   male  0.83     0     2        248738 29.0000  <NA>        S
10         120        0      3       Andersson, Miss. Ellis Anna Maria female  2.00     4     2        347082 31.2750  <NA>        S
# ... with 52 more rows
>
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com

EXPLORATORY DATA ANALYSIS

MANGOSOLUTIONS
data analysis that delivers

Nic Crane – Data Scientist
ncrane@mango-solutions.com

EDA - What is EDA?

- First look at data
- Helps form hypotheses and assess assumptions
- Graphical and numerical

Nic Crane – Data Scientist
ncrane@mango-solutions.com

EDA - Anscombe's Quartet

Dataset 1

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Dataset 2

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

Dataset 3

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Dataset 4

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

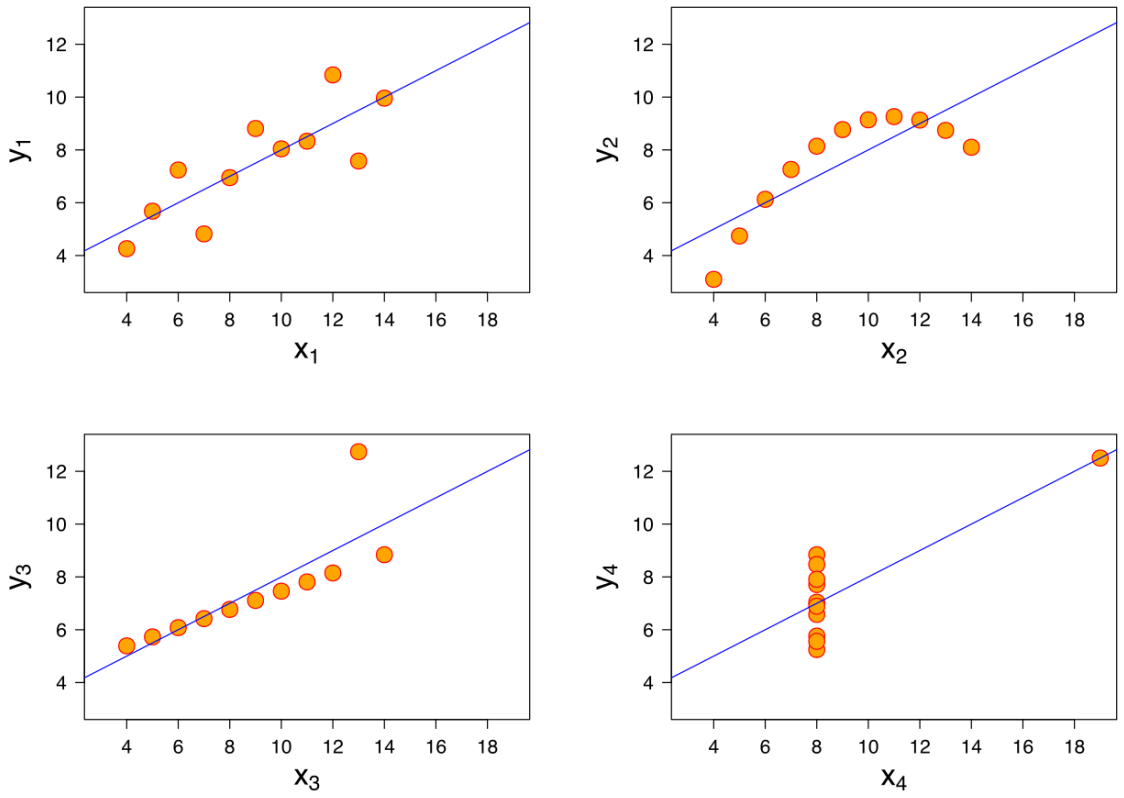
Nic Crane – Data Scientist
ncrane@mango-solutions.com

EDA - Anscombe's Quartet

Property	Value
Mean of x in each case	9 (exact)
Variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

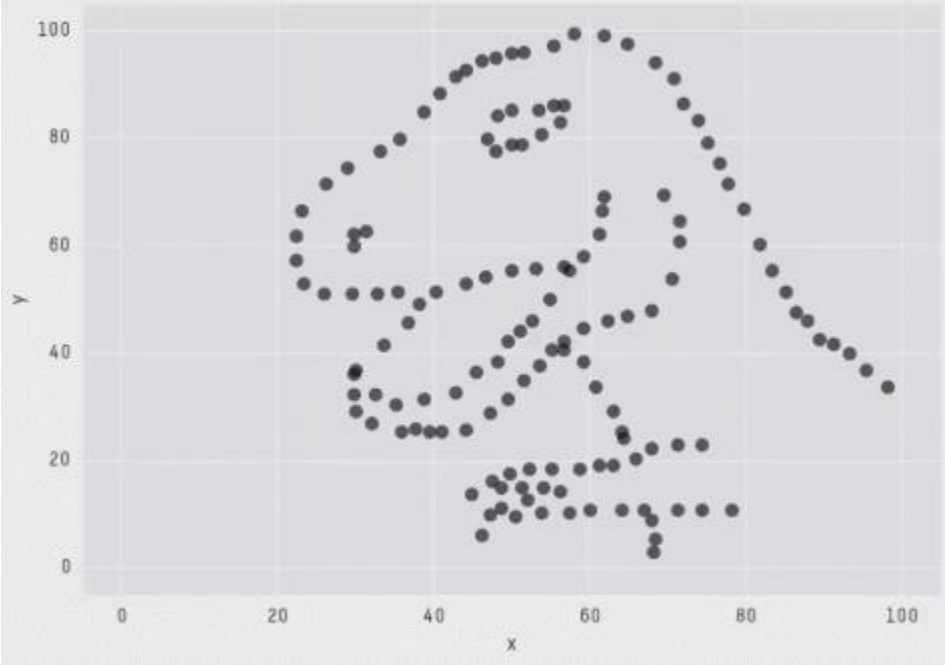
Nic Crane – Data Scientist
ncrane@mango-solutions.com

EDA - Anscombe's Quartet



Nic Crane – Data Scientist
ncrane@mango-solutions.com

EDA - Datasaurus Dozen



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Nic Crane – Data Scientist
ncrane@mango-solutions.com

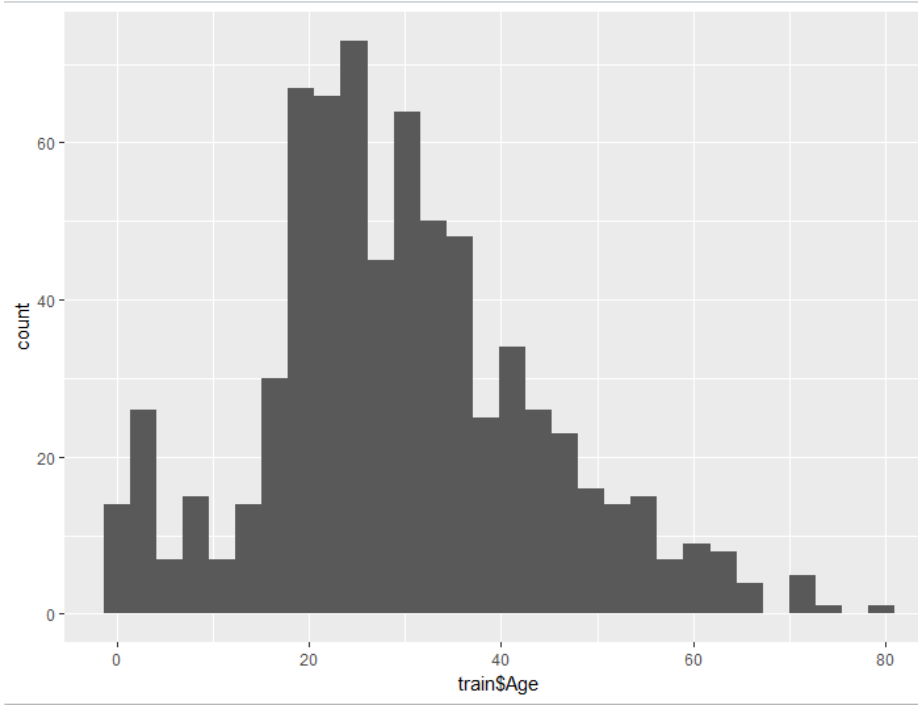
EDA - EDA in R

- Base R – fine
  - pairs
  - summary
- ggplot2 – better!
  - qplot

Nic Crane – Data Scientist
ncrane@mango-solutions.com

EDA with qplot example

```
> library(ggplot2)
> qplot(train$Age)
```



Nic Crane – Data Scientist
ncrane@mango-solutions.com

CLEANING DATA

MANGOSOLUTIONS
data analysis that delivers

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Cleaning data

- Common cleaning tasks:
  - 1 column = 1 variable
  - Missing values
  - Outliers

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Each column a variable

- Sometimes columns contain multiple variables of useful information.
- For example

```
                                                  NameSex
                                     Braund, Mr. Owen Harris_male
Cumings, Mrs. John Bradley (Florence Briggs Thayer)_female
                                     Heikkinen, Miss. Laina_female
Futrelle, Mrs. Jacques Heath (Lily May Peel)_female
                                     Allen, Mr. William Henry_male
                                             Moran, Mr. James_male
```

- This column contains information on name & gender. We should separate them.

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Splitting columns

- Columns with multiple variables are usually strings.
- grep {base} family useful for finding substrings
- strsplit {base} can separate strings based on a substring
- Regex can be used for more complex cases

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Using strsplit

- Our column:

```
                                                                NameSex
                                                 Braund, Mr. Owen Harris_male
                                 Cumings, Mrs. John Bradley (Florence Briggs Thayer)_female
                                                  Heikkinen, Miss. Laina_female
                                       Futrelle, Mrs. Jacques Heath (Lily May Peel)_female
                                                  Allen, Mr. William Henry_male
                                                        Moran, Mr. James_male
```

- The variables are split by "_"

```
# Splits column into a list
split_col <- strsplit(NameSex, split = "_")
# Extract first sub-element of each list element
name_col <- vapply(split_col, "[[", 1, FUN.VALUE = "character")
# Extract second sub-element of each list element
gender_col <- vapply(split_col, "[[", 2, FUN.VALUE = "character")
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com


MISSING VALUES

Missing values

- How do missing values affect our data?

- Missing values can affect certain algorithms and functions that aren't well equipped to deal with them.

```
# Doesnt cope with NA
mean(x)

# Copes with NA
mean(x, na.rm = TRUE)
```

- Abundance of missing values can lead to uninteresting variables and/or observations.

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Dealing with missing data

- Imagine our data looks like this

var1	var2	var3	var4
	0.608439	0.056307	0.581596
0.902304	0.761943		0.843147
0.766863	0.716702	0.682112	
0.729033	0.878197	0.469894	0.14021
0.955425		0.821906	0.612473
0.403717	0.181601	0.860137	0.349048

- Any function on this dataset that uses na.omit() will remove all but 2 rows.

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Dealing with missing data

- In this case we may want to impute values in order to keep our rows

- Common technique is the mean imputation that will replace the missing value with the column mean

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Dealing with missing data

- In this case

var1	var2	var3	var4
0.608439	0.608439	0.056307	0.581596
0.902304			
0.766863	0.716702	0.682112	0.682112
0.729033	0.878197	0.469894	0.14021
0.955425		0.821906	0.612473
0.403717	0.181601	0.860137	0.349048

- We could justify imputing the blue square but omitting the row with red squares.

Nic Crane – Data Scientist
ncrane@mango-solutions.com

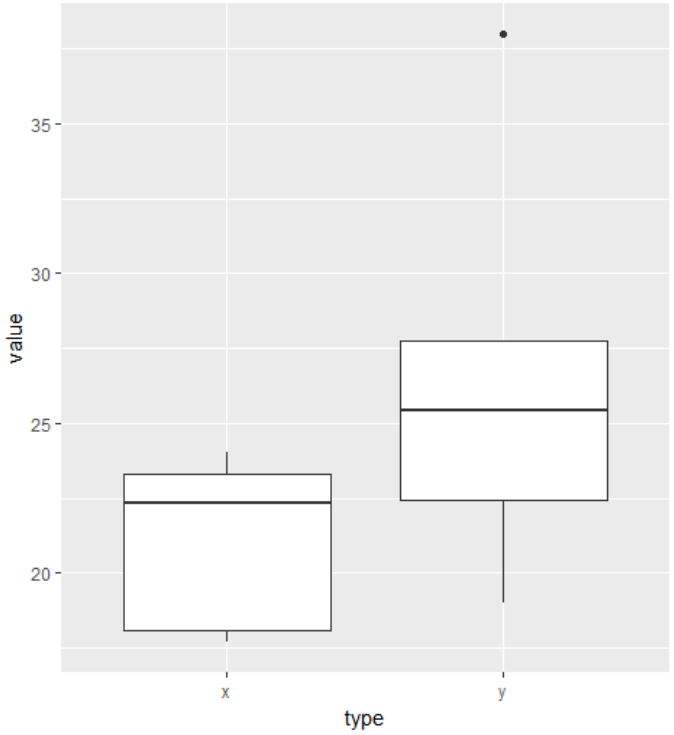
Outliers

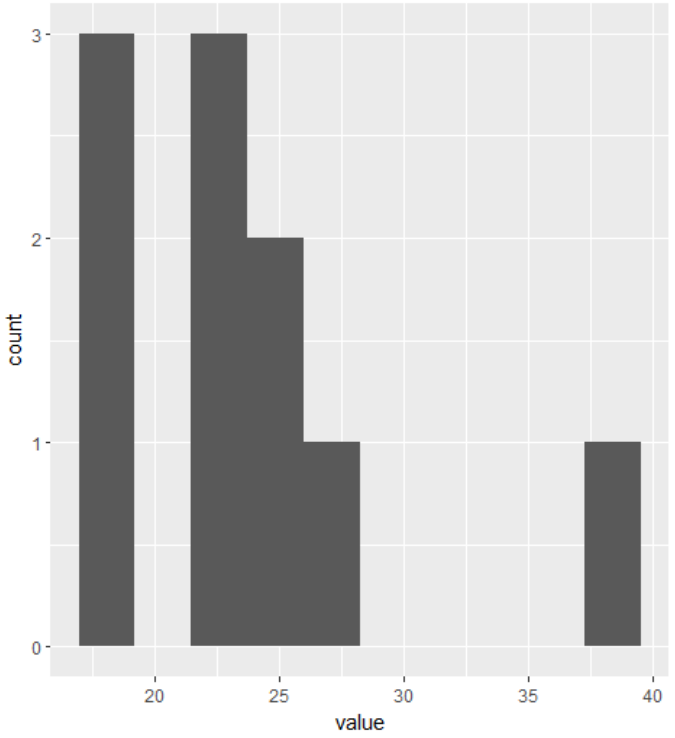
- Outliers are non-ordinary observations that appear different from the rest of the data
- They can distort statistics and be highly influential in models
- Defining what makes an outlier is quite subjective

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Finding outliers

- Boxplots and histograms

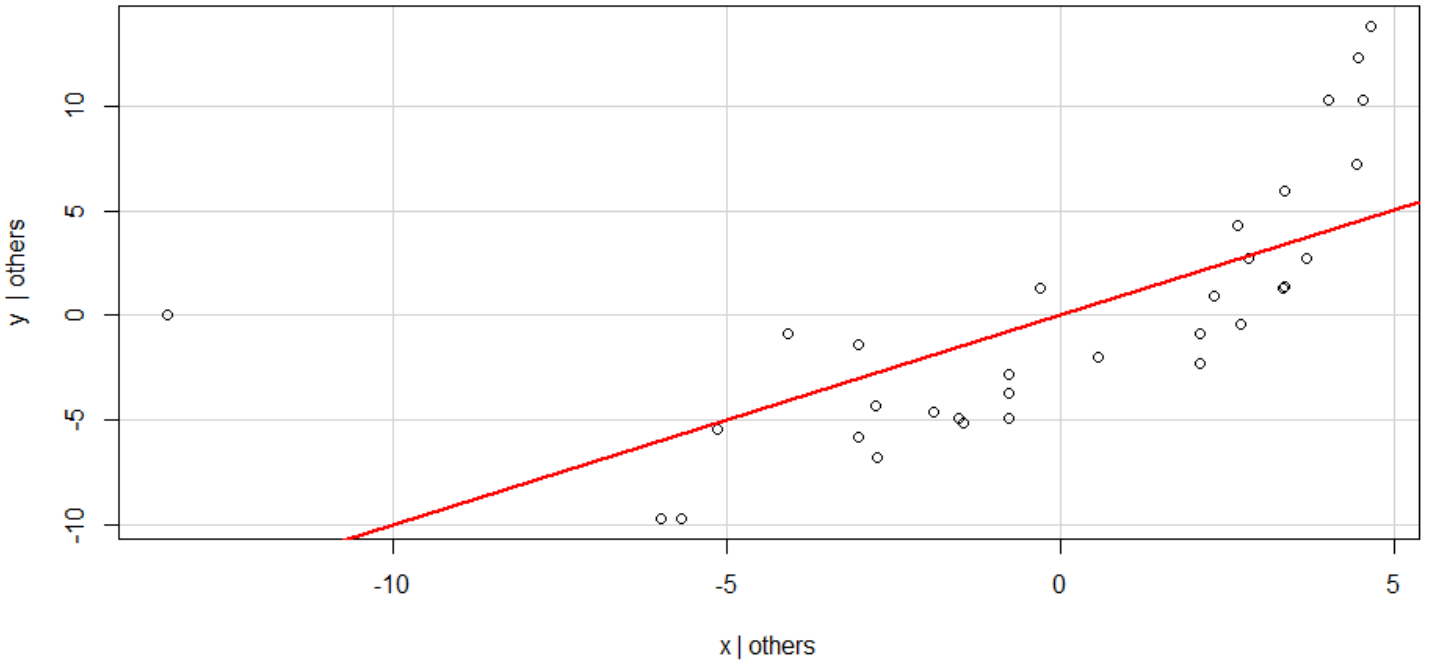




Nic Crane – Data Scientist
ncrane@mango-solutions.com

Finding outliers

- Leverage model plots



Nic Crane – Data Scientist
ncrane@mango-solutions.com

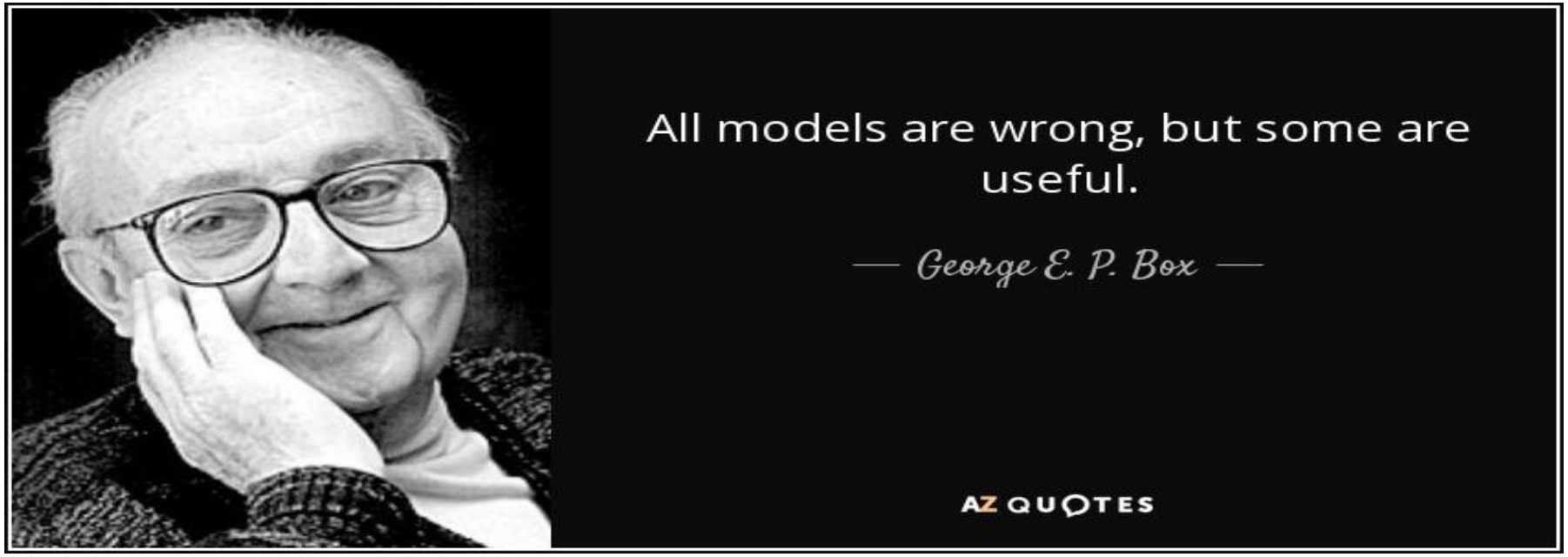
Dealing with outliers

- As a general rule, observations should not be removed unless there is *good* reasons to do so.

- Knowledge of the data is the most powerful tool to identify and remove outliers. Sometimes freakish results are an inherent part of the data structure!

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Modelling



Nic Crane – Data Scientist
ncrane@mango-solutions.com

What is a model?

- Formally, a mathematical representation of a process typically built on existing data
- Informally, it is a “rule of thumb”

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Machine learning

- So far we have formatted, explored and cleaned our data. Now we can think about creating some really valuable insights from the data.
- To do this we can use machine learning techniques.
- R has a huge variety of packages suited to machine learning.

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Unsupervised machine learning

- Unsupervised techniques are when we **don't** know the true answer.
- Our aim is to create an outcome
- Common unsupervised techniques:
  - Clustering (kmeans, hierarchical,…)
  - PCA

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Supervised machine learning

- Supervised techniques are when we create an algorithm or a model with full knowledge of the correct outcome.
- We then take this learning and apply it to data where we don't know the outcome.
- Common supervised techniques:
  - Regression
  - Classification
  - Neural networks, random forests, …

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Predicting deaths on the titanic

- Our aim is to predict which passengers died on the titanic.
- The variables in our data are:

```
PassengerId    Survived      Pclass        Name         Sex         Age
  "integer"   "integer"   "integer" "character" "character"   "numeric"
       SibSp       Parch      Ticket        Fare       Cabin    Embarked
  "integer"   "integer" "character"   "numeric" "character" "character"
```

- Where ‘Survived’ is our response variable.

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Workflow

- Split our data into test & training sets
- ‘train’ our model on the training set
- ‘test’ our model’s performance on the test set

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Splitting our data

- One of the simplest but nonetheless popular cross validation techniques is to partition the data

```
Ind <- sample(x = 1:891, size = 91)

titanic_test <- titanic[Ind,]
titanic_train <- titanic[-Ind,]
```

- Bootstrapping this process is also effective

Nic Crane – Data Scientist
ncrane@mango-solutions.com

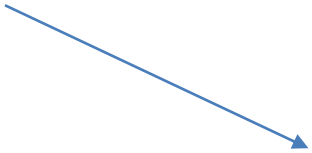
Training the model

- What do our variables look like?

```
> as.data.frame(head(titanic_train))
  PassengerId Survived Pclass                                                 Name    Sex Age SibSp Parch           Ticket    Fare Cabin Embarked
1           1        0      3                              Braund, Mr. Owen Harris   male  22     1     0        A/5 21171  7.2500  <NA>        S
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0         PC 17599 71.2833   C85        C
3           3        1      3                               Heikkinen, Miss. Laina female  26     0     0 STON/O2. 3101282  7.9250  <NA>        S
4           4        1      1         Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0           113803 53.1000  C123        S
5           5        0      3                             Allen, Mr. William Henry   male  35     0     0           373450  8.0500  <NA>        S
6           6        0      3                                     Moran, Mr. James   male  NA     0     0           330877  8.4583  <NA>        Q
```

- Preparing the data

```
train_naiive <- titanic_train %>%
  mutate(survive = factor(Survived)) %>%
  select(-c(PassengerId, Name, Survived, Ticket))
```



```
  Pclass    Sex Age SibSp Parch    Fare Cabin Embarked survive
1      3   male  22     1     0  7.2500  <NA>        S       0
2      1 female  38     1     0 71.2833   C85        C       1
3      3 female  26     0     0  7.9250  <NA>        S       1
4      1 female  35     1     0 53.1000  C123        S       1
5      3   male  35     0     0  8.0500  <NA>        S       0
6      3   male  NA     0     0  8.4583  <NA>        Q       0
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com

Random Forest model

```
model1 <- train(survive~., data = train_naiive,
                method = "rf", na.action = na.omit)
```

```
Call:
 randomForest(x = x, y = y, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 124

        OOB estimate of  error rate: 27.92%
Confusion matrix:
   0  1 class.error
0 32 19   0.3725490
1 24 79   0.2330097
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com

Improving our model

- One of our variables ‘Cabin’ is very close to a unique id. Not many passengers will share a cabin.
- What might be more useful is the cabin floor rather than the exact cabin

```
train_cfloor <- titanic_train %>%
  mutate(survive = factor(Survived),
         cabinFloor = gsub('[0-9]+', '', Cabin) %>%
           substring(1,1)) %>%
  select(-c(PassengerId, Name, Survived, Ticket, Cabin))
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com

Re-running the model

- We have some very slight improvement

```
Call:
 randomForest(x = x, y = y, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 15

        OOB estimate of  error rate: 25.97%
Confusion matrix:
   0  1 class.error
0 31 20   0.3921569
1 20 83   0.1941748
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com

Removing sparse variables

- We can calculate the percentage of missing elements for each variable:

Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	survive	cabinFloor
0.00000	0.00000	0.20625	0.00000	0.00000	0.00000	0.00125	0.00000	0.78375

- The variable cabinFloor is actually removing 78% of ours observations due to the function's inability to deal with missing rows

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Re-running without cabinFloor

```
train_noCabin <- train_cfloor %>%
  select(-cabinFloor)

model4 <- train(survive~., data = train_noCabin,
                method = "rf", na.action = na.omit)
```

```
Call:
 randomForest(x = x, y = y, mtry = param$mtry)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 18.14%
Confusion matrix:
    0   1 class.error
0 348  33  0.08661417
1  82 171  0.32411067
```


Nic Crane – Data Scientist
ncrane@mango-solutions.com

Testing our model

- We now take our trained model and apply it to the test set

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 33  8
         1  1 26

               Accuracy : 0.8676
                 95% CI : (0.7636, 0.9377)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.957e-10
```

- Not too bad!

Nic Crane – Data Scientist
ncrane@mango-solutions.com

The caret package

- Huge suite of machine learning techniques (>200!)
- A simple syntax and functions that handle cross validation needs make it an ideal one-stop shop
- Plenty of vignettes and tutorials

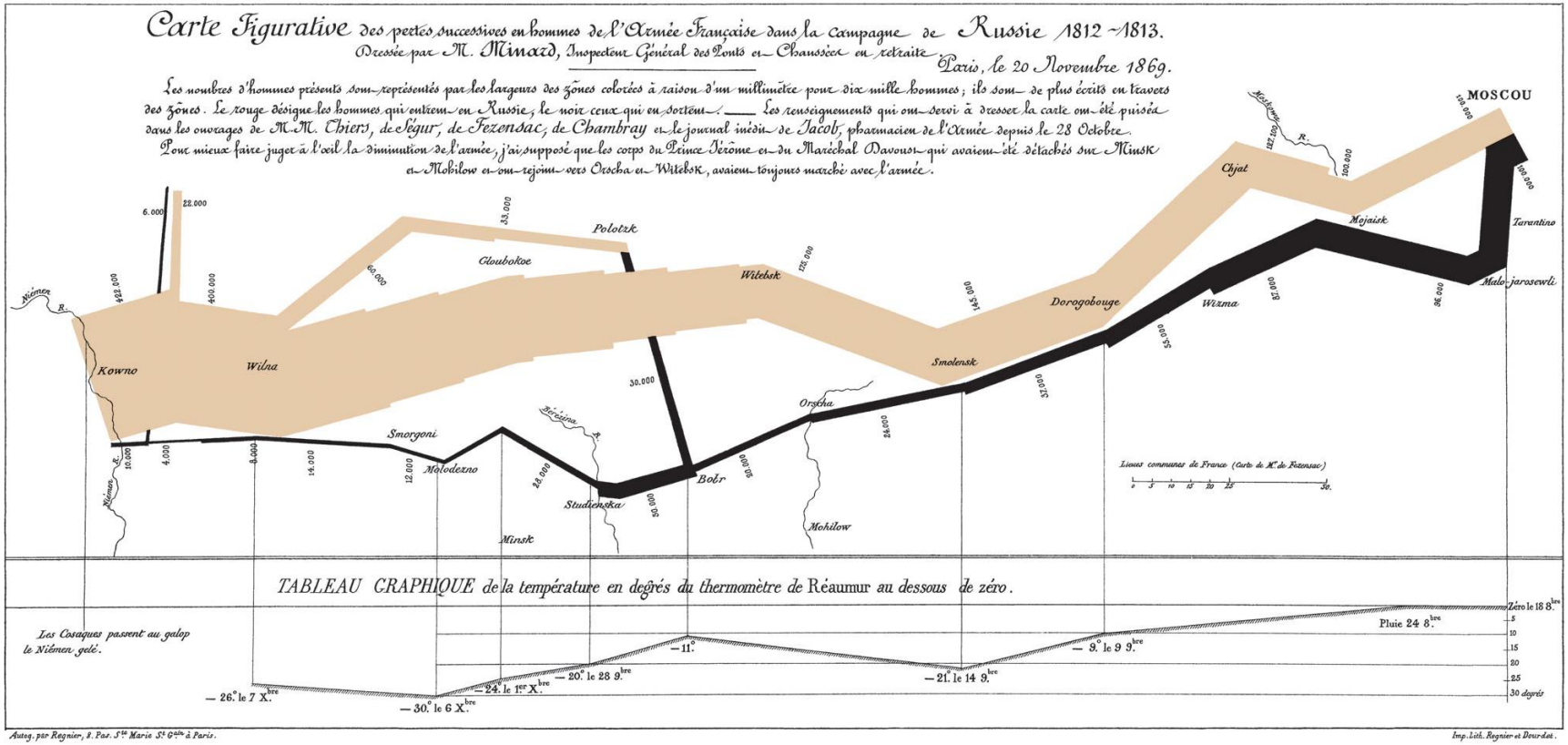
Nic Crane – Data Scientist
ncrane@mango-solutions.com

COMMUNICATING RESULTS

MANGOSOLUTIONS
data analysis that delivers

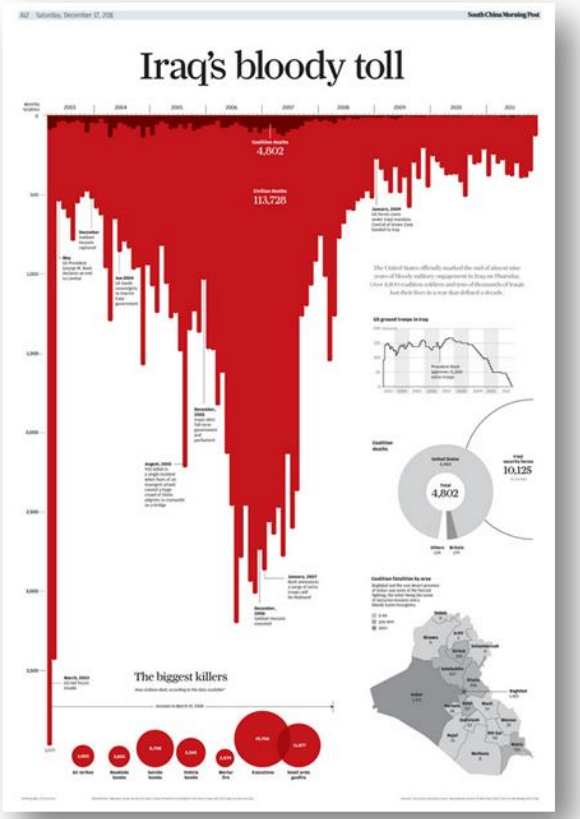
Nic Crane – Data Scientist
ncrane@mango-solutions.com

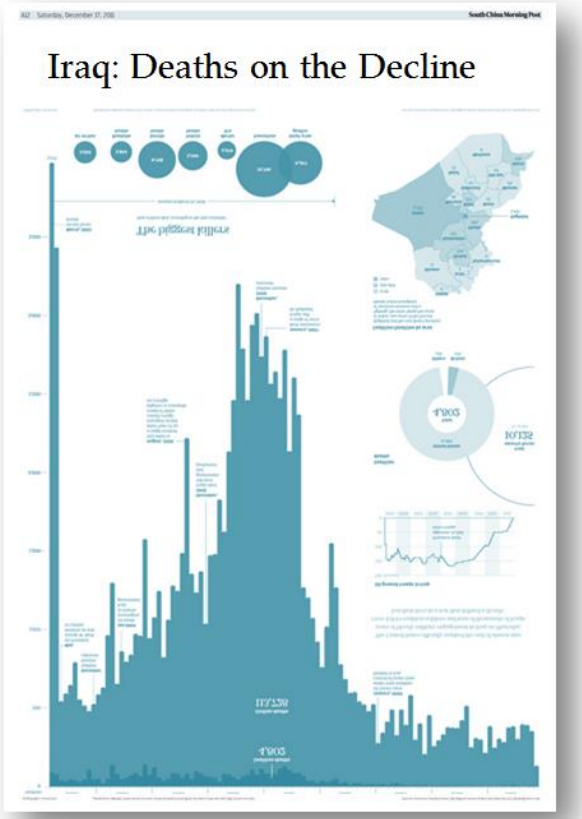
Graphics



Nic Crane – Data Scientist
ncrane@mango-solutions.com

Graphics





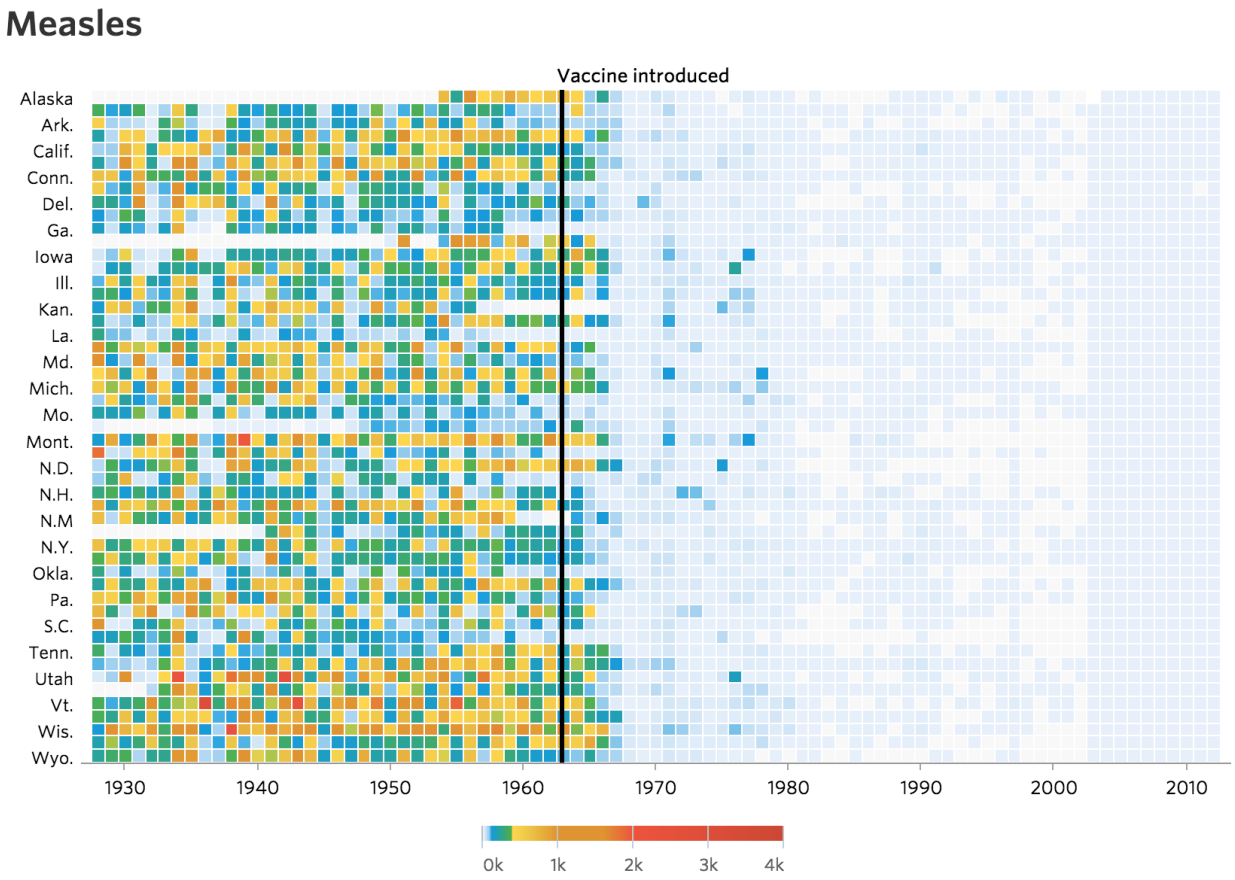
Nic Crane – Data Scientist
ncrane@mango-solutions.com

Graphics in R?

- Base
- lattice
- **ggplot2**

Nic Crane – Data Scientist
ncrane@mango-solutions.com

ggplot2



MANGOSOLUTIONS
data analysis that delivers

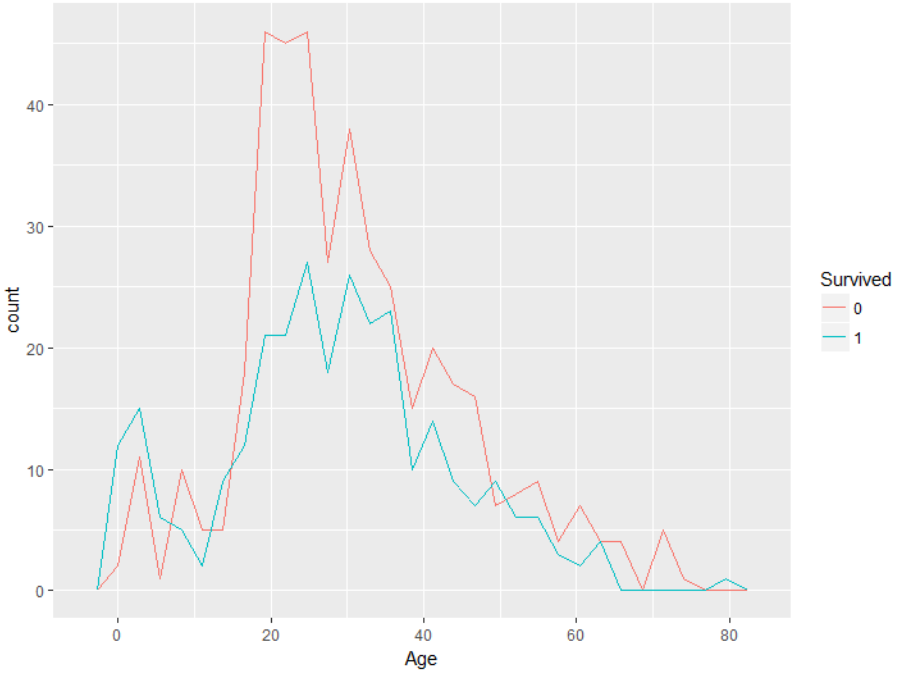
Nic Crane – Data Scientist
ncrane@mango-solutions.com

ggplot2

Nic Crane – Data Scientist
ncrane@mango-solutions.com

ggplot2 Titanic Example

```
> train$Survived <- as.factor(train$Survived)
> qplot(Age,data=train,colour=Survived,geom="freqpoly")
```



Nic Crane – Data Scientist
ncrane@mango-solutions.com

R Markdown

- Markdown = simple markup language
- Embed chunks of R code
- Reproducible
- Output to documents, slides, books, websites

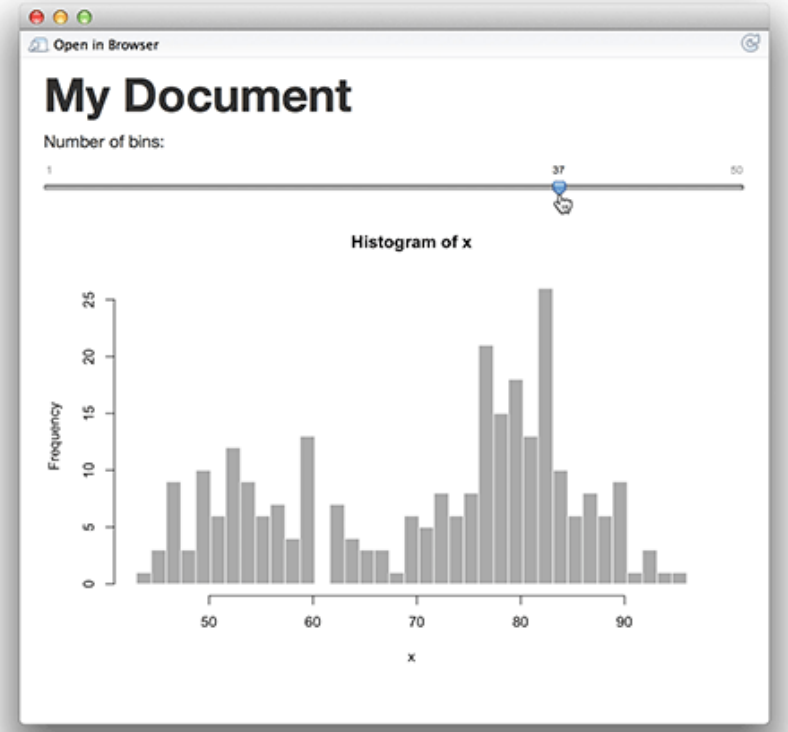
Nic Crane – Data Scientist
ncrane@mango-solutions.com

Shiny

- Build web applications directly from R
- Uses HTML, CSS and JavaScript – but you don't need to know any!
- Hosting available on shinyapps.io

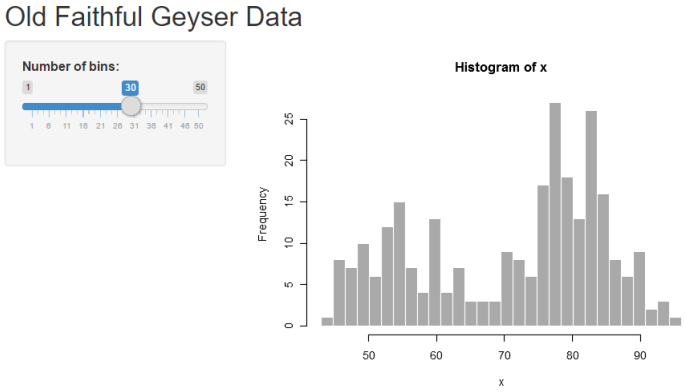
Nic Crane – Data Scientist
ncrane@mango-solutions.com

Shiny



Nic Crane – Data Scientist
ncrane@mango-solutions.com

Shiny




```
#
# This is a Shiny web application. You can run the application by clicking
# the 'Run App' button above.
#
# Find out more about building applications with Shiny here:
#
#    http://shiny.rstudio.com/
#

library(shiny)

# Define UI for application that draws a histogram
ui <- fluidPage(

   # Application title
   titlePanel("Old Faithful Geyser Data"),

   # Sidebar with a slider input for number of bins
   sidebarLayout(
      sidebarPanel(
         sliderInput("bins",
                     "Number of bins:",
                     min = 1,
                     max = 50,
                     value = 30)
      ),

      # Show a plot of the generated distribution
      mainPanel(
         plotOutput("distPlot")
      )
   )
)

# Define server logic required to draw a histogram
server <- function(input, output) {

   output$distPlot <- renderPlot({
      # generate bins based on input$bins from ui.R
      x    <- faithful[, 2]
      bins <- seq(min(x), max(x), length.out = input$bins + 1)

      # draw the histogram with the specified number of bins
      hist(x, breaks = bins, col = 'darkgray', border = 'white')
   })
}

# Run the application
shinyApp(ui = ui, server = server)
```

MANGOSOLUTIONS
data analysis that delivers

Nic Crane – Data Scientist
ncrane@mango-solutions.com

Shiny App Titanic Example

https://attbigdatagroup.shinyapps.io/Titanic-Shiny-Application/

Nic Crane – Data Scientist
ncrane@mango-solutions.com

PACKAGING AND CODE STRUCTURE

MANGOSOLUTIONS
data analysis that delivers

Nic Crane – Data Scientist
ncrane@mango-solutions.com

R Packages

- Collection of code, documentation, data, with a pre-specified structure
- Easily shareable code
- Simplify loading of code and packages
- Maintain a single version and be able to identify which version of code is being used
- Provide documentation and usage examples of code easily

Nic Crane – Data Scientist
ncrane@mango-solutions.com

R Packages

- R Packages by Hadley Wickham
- Book available at: http://r-pkgs.had.co.nz/

Nic Crane – Data Scientist
ncrane@mango-solutions.com