MANGO SOLUTIONS

Programming with dplyr

Adnan Fiaz

Data Scientist

afiaz@mango-solutions.com

@tapundemek

Introduction: who am I?

- Data Scientist @ Mango Solutions (previously KLM)
- Background in business maths
- useR since +/- 2012

Introduction: who is Mango?

- Data Science consultancy
- Offices in Chippenham & London
- +/ 25 Data Scientists and growing...
- Clients include M&S , S&P, ONS, BCA and many more acronyms
- Organisers of EARL


EARL
Exeter R 10% DISCOUNT:
RUSERSRCOOL10
www.earlconf.com

Agenda

- Refresh
- Writing Functions
- Bang! Bang!
- Quote first, shoot later

A Grammar of Data Manipulation

- Consistent and fast way
- Manipulate data through simple “verbs”
- Abstract away from backend
  - Database
  - data.table

str(dplyr)

Basic : (filter, select, mutate, arrange, slice, pull, rename)

Select: (starts_with, ends_with, contains, one_of, everything, matches)

2-table: (left_join, right_join, full_join, semi_join, anti_join, intersect, setdiff, union)

Util: combine, bind_rows, bind_cols, between, glimpse, n, row_number, rowwise, sample_n, top_n, tally

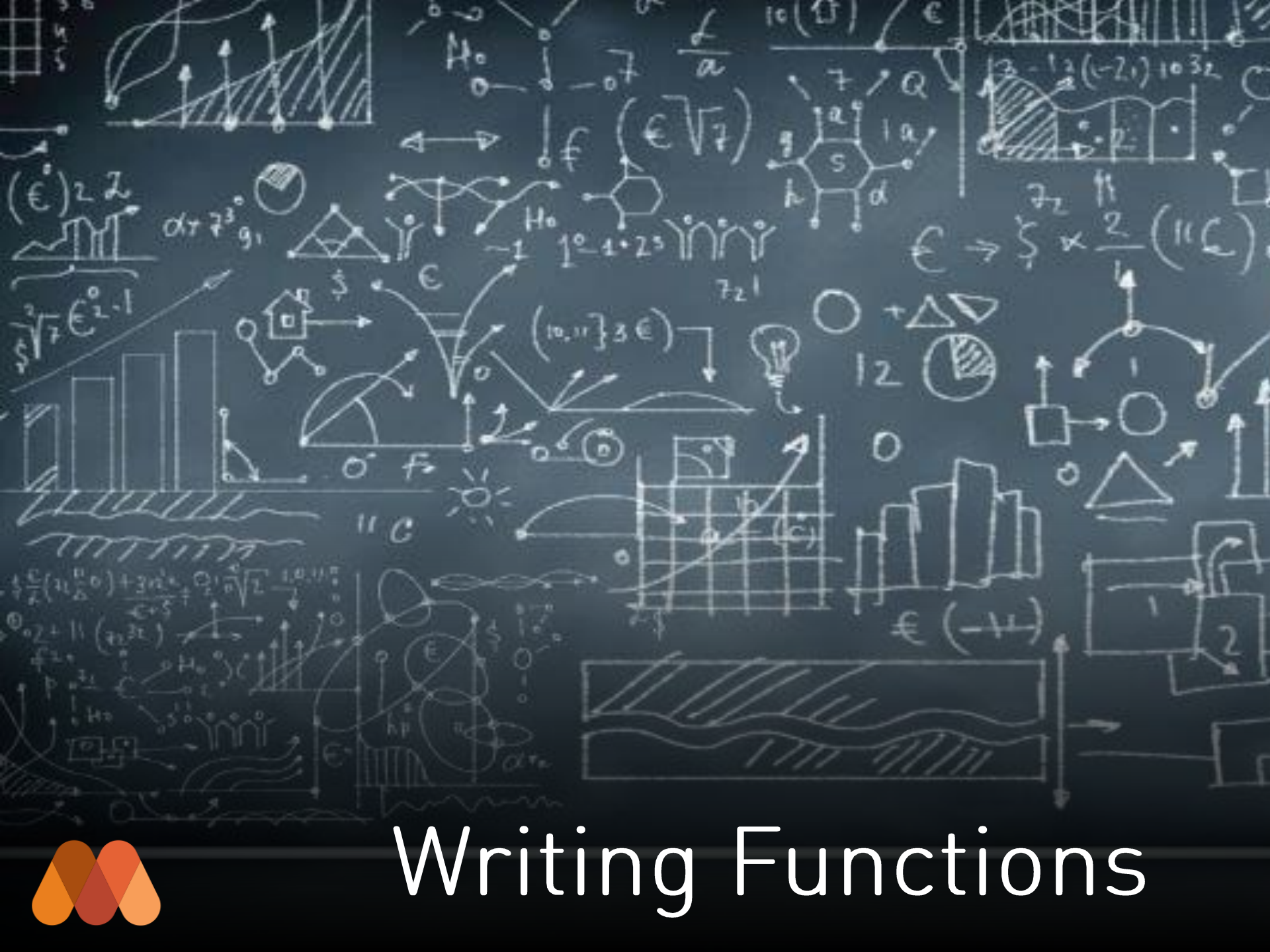
Window: lead, lag, cumall, cumany, cumsum, cummean, min_rank, percent_rank

SAC: groupby, do, summarise, count

Example

```
tubeData %>%
  group_by(Line) %>%
  summarise(mean = mean(Excess))
```

```
## # A tibble: 10 x 2
##                Line     mean
##              <fctr>    <dbl>
##  1         Bakerloo 5.047714
##  2          Central 5.998667
##  3    Circle & Ham 7.166095
##  4         District 5.485619
##  5          Jubilee 5.809238
##  6    Metropolitan 8.553048
##  7         Northern 5.714095
##  8      Piccadilly 5.942095
##  9         Victoria 5.914190
## 10 Waterloo & City 2.058381
```



Writing Functions

Example (cont.)

- Replace the group_by input by a variable
- …within a function

```
excess <- function(groupVar){
  tubeData %>%
    group_by(groupVar) %>%
    summarise(mean = mean(Excess))
}
excess("Line")
# Error in grouped_df_impl(data, unname(vars), drop) : Column `groupVar` is unknown
```

Why does it fail?

- Remember base R?
  - `df[, "column"]`
  - `df[df$column == x, ]`
- This is what dplyr simplifies for you
- But we don't want that

- We need to tell dplyr verbs that (forcefully)

Bang! Bang!

The !! (bang bang) operator

- The !! evaluates the variable
- The value is then passed on to dplyr verbs

Example (cont.)

```
excess <- function(groupVar){
  tubeData %>%
    group_by(!!groupVar) %>%
    summarise(mean = mean(Excess))
}
excess("Line")
```

```
## # A tibble: 1 x 2
##   `"Line"`     mean
##      <chr>    <dbl>
## 1     Line 5.768914
```

Why does it still fail?

- The variable is evaluated within the context of the data
- We want it to evaluate within the context it was defined in


Quote first, shoot later

Quosures

- Capture environment of creation
- Don’t evaluate expressions
- Quosures: quoting enclosures

Example (cont.)

```
excess <- function(groupVar){
  quoVar <- enquo(groupVar)
  print(quoVar)
  tubeData %>%
    group_by(!!quoVar) %>%
    summarise(mean = mean(Excess))
}
excess(Line)
```

```
## <quosure: global>
## ~Line
```

```
## # A tibble: 10 x 2
##                Line     mean
##              <fctr>    <dbl>
##  1        Bakerloo 5.047714
##  2         Central 5.998667
##  3    Circle & Ham 7.166095
##  4        District 5.485619
##  5         Jubilee 5.809238
##  6    Metropolitan 8.553048
##  7        Northern 5.714095
##  8      Piccadilly 5.942095
##  9        Victoria 5.914190
## 10 Waterloo & City 2.058381
```

More advanced...

```
excess <- function(groupVar, summariseVar){
  qGroupVar <- enquo(groupVar)
  qSummariseVar <- enquo(summariseVar)

  tubeData %>%
    group_by(!!qGroupVar) %>%
    summarise(mean = mean(!!qSummariseVar))
}

excess(Line, Excess*10)
```

More advanced...

```
excess <- function(groupVar, summariseVar){
  qGroupVar <- enquo(groupVar)
  qSummariseVar <- enquo(summariseVar)
  resultName <- paste0("mean_", quo_name(qSummariseVar))

  tubeData %>%
    group_by(!!qGroupVar) %>%
    summarise(!!resultName := mean(!!qSummariseVar))
}

excess(Line, Excess)
```

HOW DARE YOU QUESTION ME!!

Adnan Fiaz

afiaz@mango-solutions.com

Links

- https://cran.r-project.org/web/packages/dplyr/vignettes/programming.html
- https://schd.ws/hosted_files/user2017/43/tidyeval-user.pdf