

Defining My EDA Project Workflow

Boston useR meetup - Nov' 2021

Priyanka Gagneja

2021-12-01

About me



- Data Science Practitioner/ Freelancer
 - Data Science Consultant with OnPoint Insights
 - TA for Applied Linear Modelling at Upenn
- Avid R user and active community member (RLadies and R4DS)
- MS, Applied Economics from Boston College
- B.Tech and MBA in India
- Women in DS, Women in Analytics Ambassador

Contact:

Twitter: @priyankaigit

Linkedin: priyanka-gagneja

Blog: <https://fun-with-data.netlify.app/>

Motivation

Doing lots of Exploratory projects

- Repetitive
- Taking so much time
 - No time for insights

Absolute first steps

Packages I discovered

Absolute first steps

- `{DataExplorer}`
 - aims to perform and simplify the 3 main goals throughout the data exploration process - Exploratory Data Analysis (EDA), Feature Engineering and Data Reporting
 - `create_report(txhousing)`
- `{DataReporter}`
 - thorough summary of the data checks and the results that make it easy to identify possible errors
 - `makeDataReport(txhousing)`
- `{skimr}`
 - provide summary statistics about variables with easy to modify defaults
 - `skim(txhousing)`

Further Exploration

Packages I discovered contd...

Further Exploration

- {rpivotTable}
- {esquisse}
- {chronicle} (Thanks to useR!2021)

{rpivotTable}

Allows

- Quick & Dirty Data Exploration

Demo

```
library(rpivotTable)

ggplot2::txhousing %>%
  rpivotTable::rpivotTable()
```

Benefits

- More than just counts, more like pivot
- Pipeable

Limitations

- No code
- Colors

{esquisse}

Allows

- Quick Exploration
- Wrangling on the go
- Code Generation

Demo

```
library(esquisse)  
esquisse::esquisser()
```

Brings up a new window, shiny-app based.

Benefits

- Code generated for each action
 - can be sent to script or console

Limitations

- Does not provide much of summaries
- Limited plot options

{chronicle}

Allows

- Quick Exploration
- Report Generation

{chronicle}

Demo

```
make_barplot <- function(dt,
  bars,
  value = NULL,
  break_bars_by = NULL,
  up_to_n_bars = 20,
  horizontal = FALSE,
  sort_by_value = horizontal,
  sort_decreasing = TRUE,
  ggtheme = 'minimal',
  x_axis_label = NULL,
  y_axis_label = NULL,
  plot_palette = NULL,
  plot_palette_generator = 'plasma',
  static = FALSE){
```

```
# create bar plot
barplot <- ggplot2::ggplot(plot_dt,
  ggplot2::aes(x = .data[[bars]],
               y = .data[[value]],
               fill = .data[[ifelse(test = is.null(break_bars_by),
                                     yes = bars,
                                     no = break_bars_by)]]) +

  ggplot2::geom_bar(stat = 'identity', alpha = .95) +
  ggtheme() +
  ggplot2::theme(panel.background = ggplot2::element_rect(fill = "transparent", colour = NA),
                 plot.background = ggplot2::element_rect(fill = "transparent", colour = NA)) +
  ggplot2::scale_y_continuous(labels = scales::number_format(accuracy = 0.01,
                                                            decimal.mark = '.',
                                                            big.mark = ',')) +

  ggplot2::scale_fill_manual(values = plot_palette)

# axes
```

{chronicle}

Demo

- add_* functions
 - make_* functions
- render_report()
 - rmarkdown::render()
- report_columns()
 - skimr::skim()

Benefits

- reduces the time to generate plots (without getting into ggplot2 layering detail)
- can generate reports in multiple formats with one rendering action

Limitations

- No ggplot code per se

Resources/References

<https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html>

<https://github.com/ekstroem/dataReporter>

<https://cran.r-project.org/web/packages/skimr/vignettes/skimr.html>

<https://github.com/smartinsightsfromdata/rpivotTable>

<https://cran.r-project.org/web/packages/esquisse/vignettes/get-started.html>

<https://github.com/pheymanss/chronicle>

For the bonus packages

<https://choonghyunryu.github.io/dlookr/articles/introduce.html>

<https://descriptr.rsquaredacademy.com/>

Bonus

- {dlookr}
- {descriptr}

Thank You !!