Actuarial Data Science: R in Insurance

Data Analytics Day

November 2019

Yuan Tian

Agenda

- Why R
- Example of using R in Insurance
  - Pricing
  - Reserving
  - Data Visualisation
  - Linking to Excel, Access Database,SQL
- Learning resources and next steps
- Hands-on session with short exercises

Why R

- Popular choice for actuaries ,scientists and academics
  - Easy and fun to learn : coding not complex
  - Already been studied and developed by many actuaries
  - Many practical package exists : ChainLadder, GoogleVis, GLM…
  - New IFOA Exam Syllabus
- Easy to share the output , e.g. writing reports, building dashboards
  - R Markdown : html, pdf, excel data…
  - R Shiny : interactive web tool
- Free!
- Data Science and Data Analytics skills :
  - MUST HAVE FOR ACTUARIES!

R vs Python

Parameter	R	Python
Objective	Data Analysis and Statistical Modeling	Data Science, Web Development, Embedded Systems
Workability	Consists of many easy to use packages	Can easily perform matrix computation as well as optimization
Integration	Locally Run Programs	Programs integrated with web-app for easy deployment
Database Handling Capacity	Poses problem for handling large dataset	Can handle large data easily without any fault
IDE	Rstudio, R GUI	Spyder, IPython, Jupyter Notebook
Essential Packages and library	ggplot2, tidyverse, caret	Numpy, pandas, scipy, scikit-learn, TensorFlow

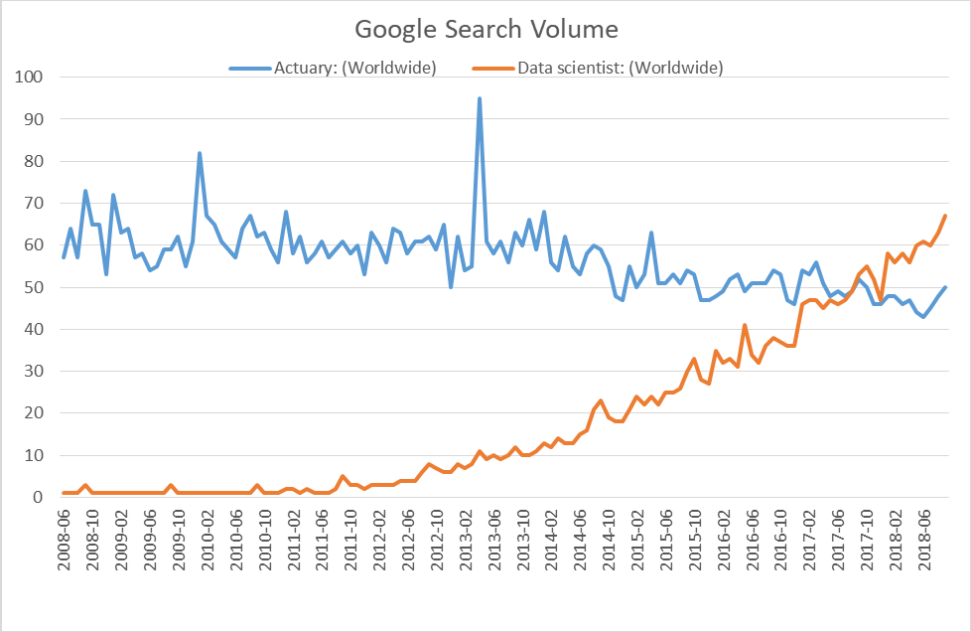
Python : multi-purpose programming language

R : statistical computing and graphical computation

Actuary vs Data Scientist

- Business Awareness, communicating insights to the business
- Core technical actuarial skills
- On going training , qualifications, CPD…

- Lack of :
  - Limited skills to deal with unstructured data
  - Large data analysis e.g web scraping
  - Visual representation ability
  - Programing skills



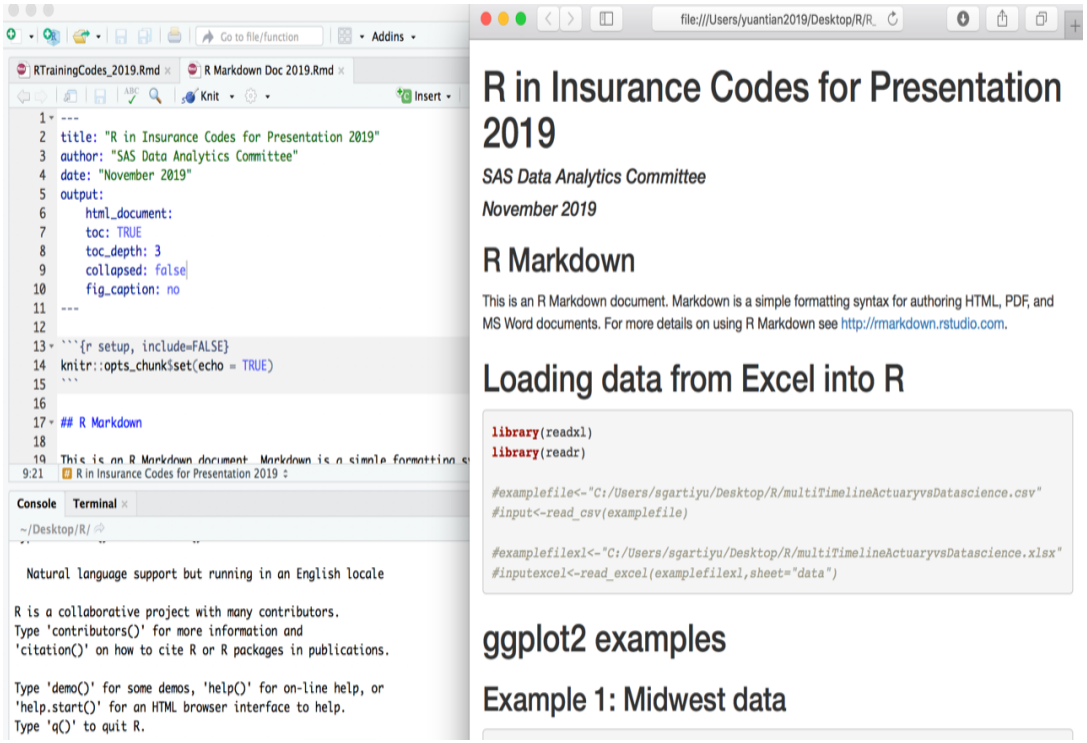
Google Trends as at Sept 2018
https://trends.google.com/trends/explore?date=all&q=Actuary,Data%20scientist

R Examples

R examples -
R Reporting RMarkdown

This Package works in Rstudio allowing to add narrative text alongside a consistent set of exhibits

- Reports have same structure every time
- Actuaries concentrate on interpretation and narrative
- Ad-hoc exhibits can be added to appendices and reproduced easily without adding workload
- Reports can be output to Word, pdf, html
- Attached in the Appendix



https://rmarkdown.rstudio.com/

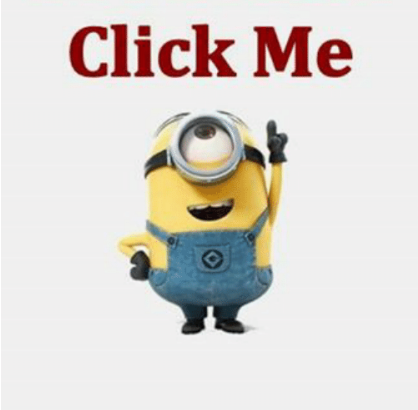
R examples -
R Reporting RMarkdown

- Open your Rstudio
- Open this html file :

  R Markdown Doc 2019.html

  RTrainingCodes_2019.html

- Try the examples!
- Let's do it together !



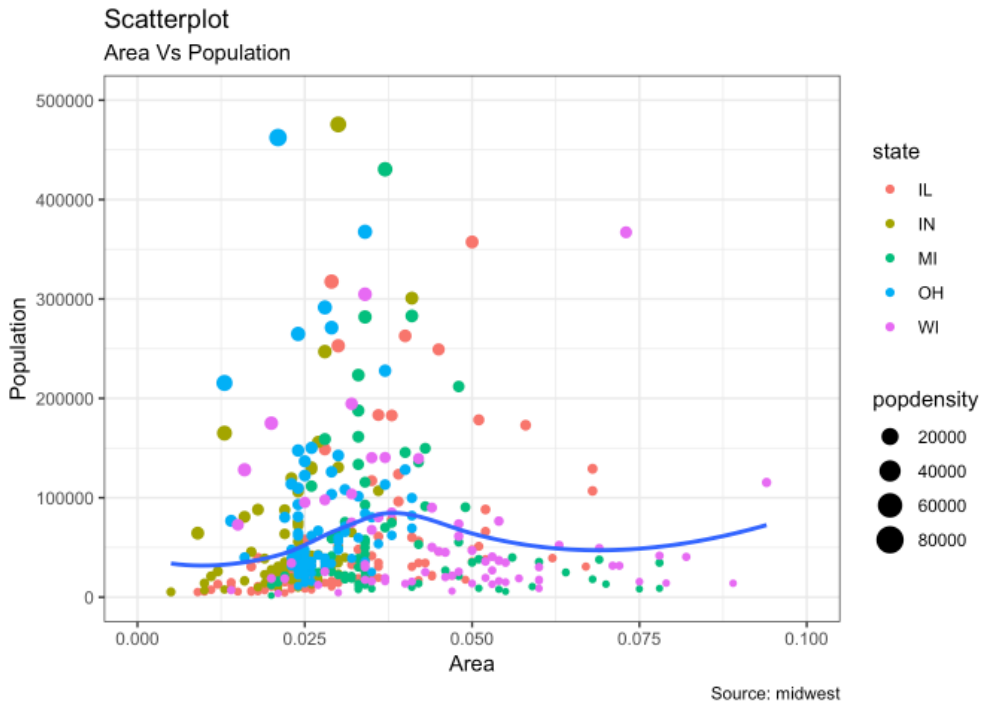
https://rmarkdown.rstudio.com/

Let’s try this example together :

ggplot2 examples

Example 1: Midwest data

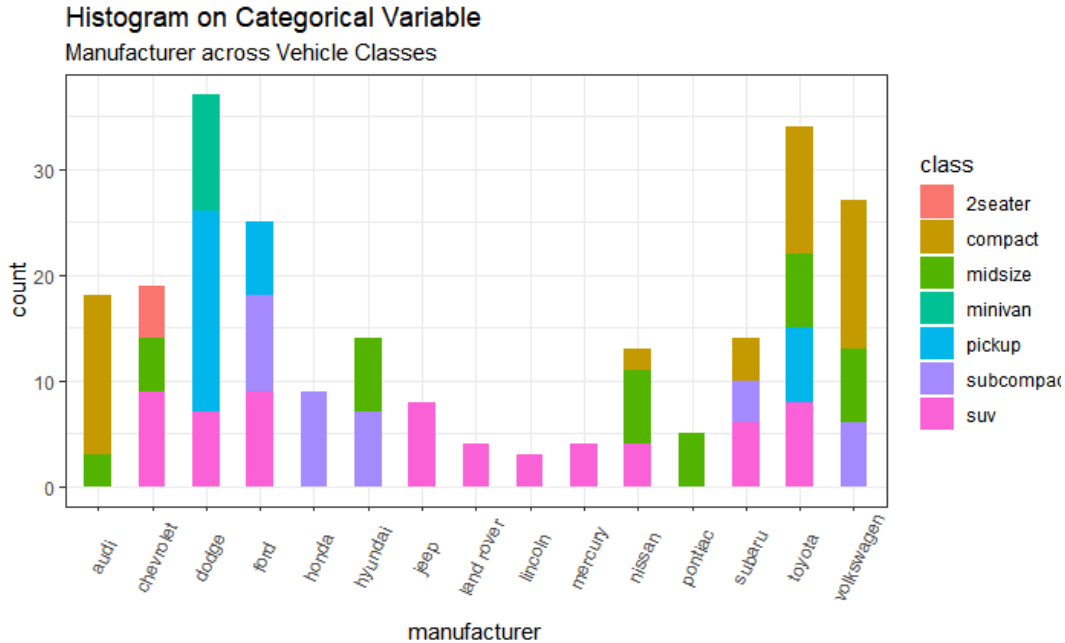
```
library(ggplot2)
midwest <- read.csv("http://goo.gl/G1K41K")
options(scipen=999)
theme_set(theme_bw())  # pre-set the bw theme.
data("midwest", package = "ggplot2")

# Scatterplot
gg <- ggplot(midwest, aes(x=area, y=poptotal)) +
  geom_point(aes(col=state, size=popdensity)) +
  geom_smooth(method="loess", se=F) +
  xlim(c(0, 0.1)) +
  ylim(c(0, 500000)) +
  labs(subtitle="Area Vs Population",
       y="Population",
       x="Area",
       title="Scatterplot",
       caption = "Source: midwest")

plot(gg)
```

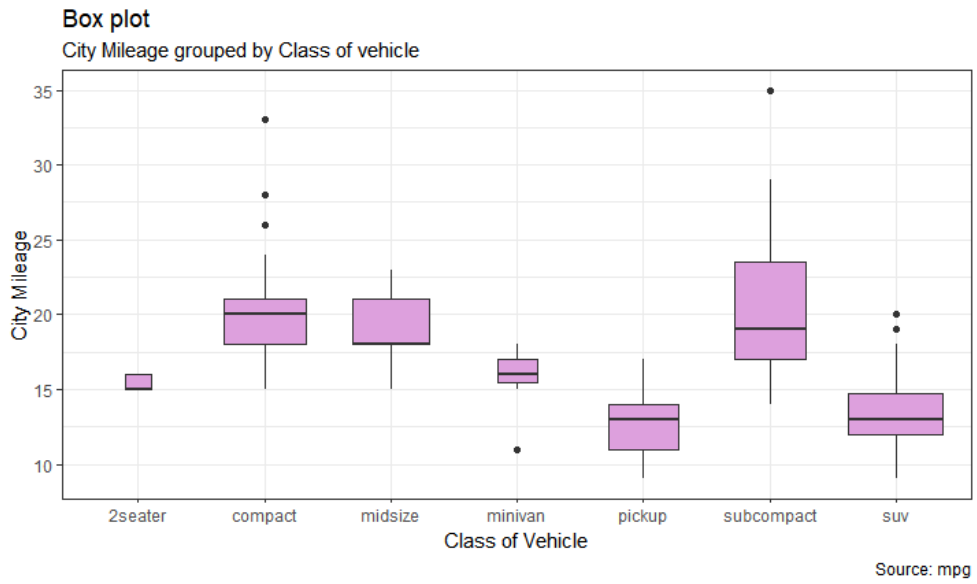


Source data and code : http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#top

R examples -
Data Visualization ggplot2

	manufactu	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
1	audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
2	audi	a4	1.8	1999	4	manual(m	f	21	29	p	compact
3	audi	a4	2	2008	4	manual(m	f	20	31	p	compact
4	audi	a4	2	2008	4	auto(av)	f	21	30	p	compact
5	audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
6	audi	a4	2.8	1999	6	manual(m	f	18	26	p	compact

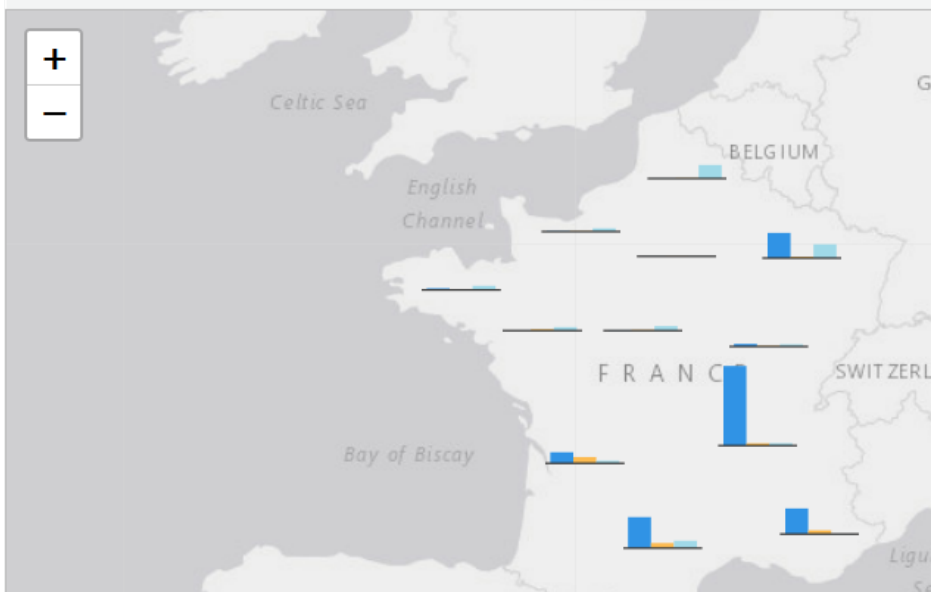


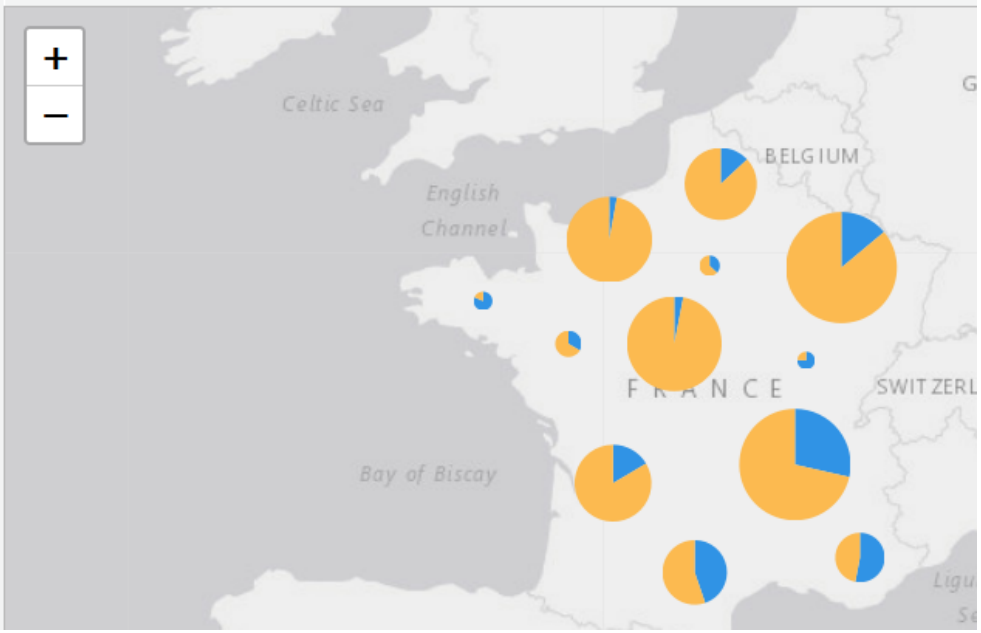


Source data and code : http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#top

R examples - Data Visualization leaflet

- Output Geographical Information to a map using *leaflet* package
- Interactive maps allow zooming and slicing
- Widely used in personal line motor pricing and Nat Cat perils aggregation management

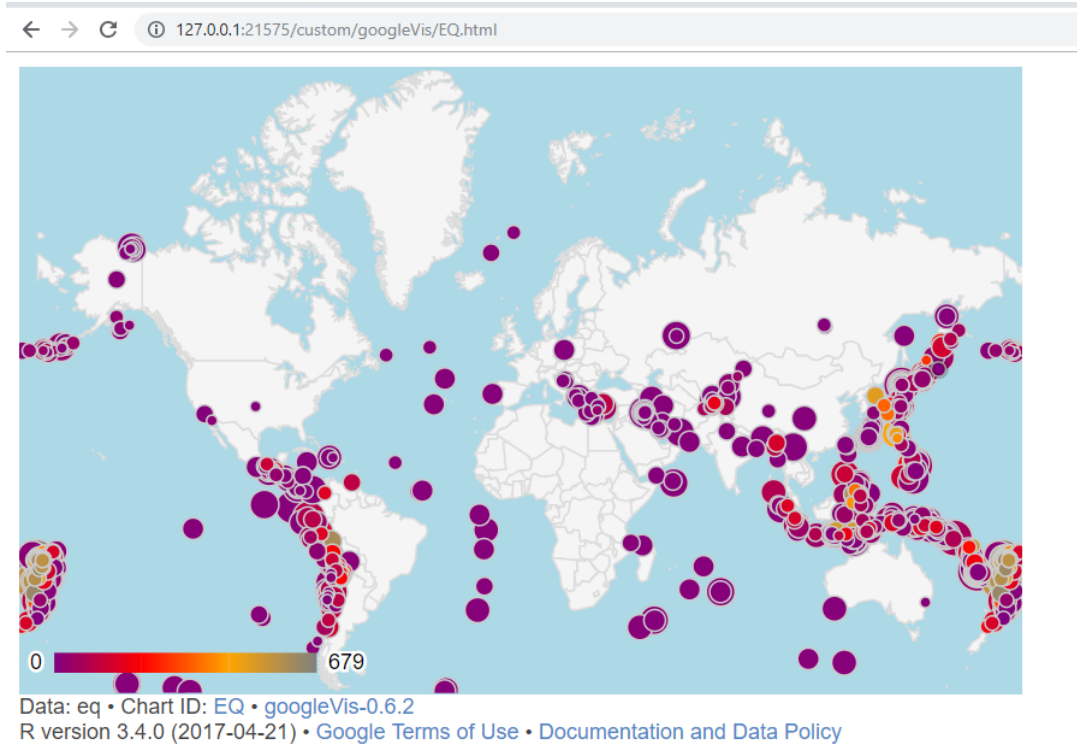




https://cran.r-project.org/web/packages/leaflet.minicharts/vignettes/introduction.html

R examples -
Data Visualization googleVis

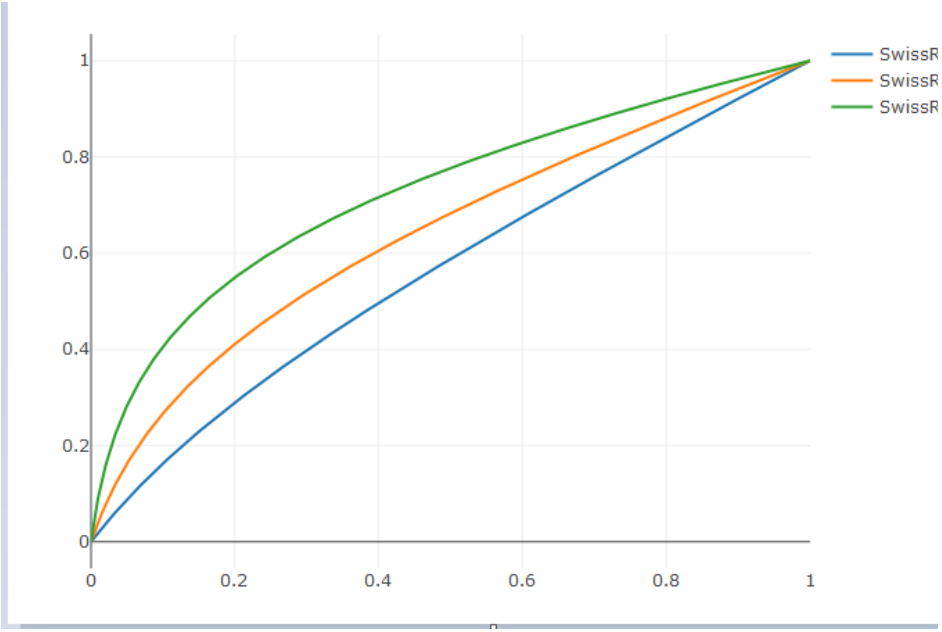
- Web interactive plot (use google chrome)
- R interface to google chart
- Author : Markus Gesmann (He is also the author of other actuarial R packages e.g. MBBEFD)

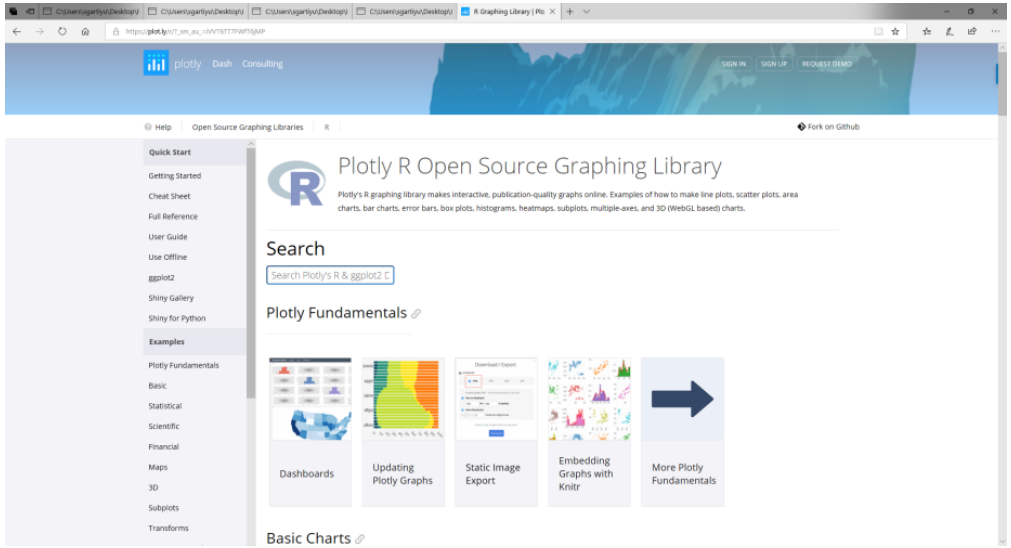


https://cran.r-project.org/web/packages/googleVis/vignettes/googleVis_examples.html

R examples - Data Visualization plotly

- Web interactive plot examples :
  - https://plot.ly/r/





https://plot.ly/ggplot2/getting-started/#plotly-for-r

R examples - Pricing

- The Basic R package : GLM
- Package MGCV : GAM
  - Generalised Additive Models
- Package XGBoost
  - Efficient linear model solver and tree learning algorithms

https://cran.r-project.org/web/packages/mgcv/mgcv.pdf
https://cran.r-project.org/web/packages/xgboost/index.html

R examples - Pricing GLM

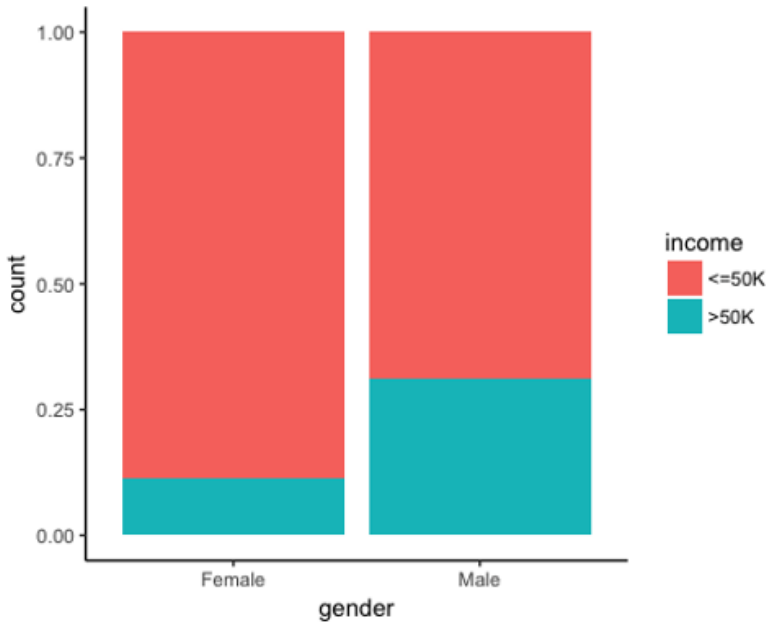
Dataset :

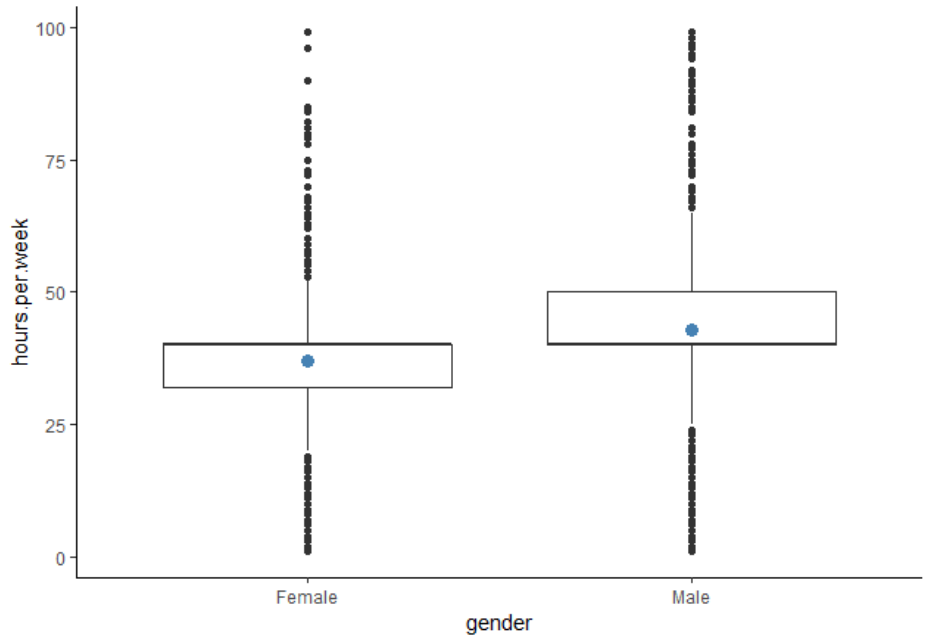
age	workclass	education	educational.num	marital.status	race	gender	hours.per.week	income
25	Private	11th	7	Never-married	Black	Male	40	<=50K
38	Private	HS-grad	9	Married-civ-spous	White	Male	50	<=50K
28	Local-gov	Assoc-acdm	12	Married-civ-spous	White	Male	40	>50K
44	Private	Some-college	10	Married-civ-spous	Black	Male	40	>50K
34	Private	10th	6	Never-married	White	Male	30	<=50K
63	Self-emp-not-inc	Prof-school	15	Married-civ-spous	White	Male	32	>50K
24	Private	Some-college	10	Never-married	White	Female	40	<=50K
55	Private	7th-8th	4	Married-civ-spous	White	Male	10	<=50K
65	Private	HS-grad	9	Married-civ-spous	White	Male	40	>50K
36	Federal-gov	Bachelors	13	Married-civ-spous	White	Male	40	<=50K

https://www.guru99.com/r-generalized-linear-model.html

R examples - Pricing GLM

Checking Data Variables e.g. factor affect earnings , boxplot, variable correlation

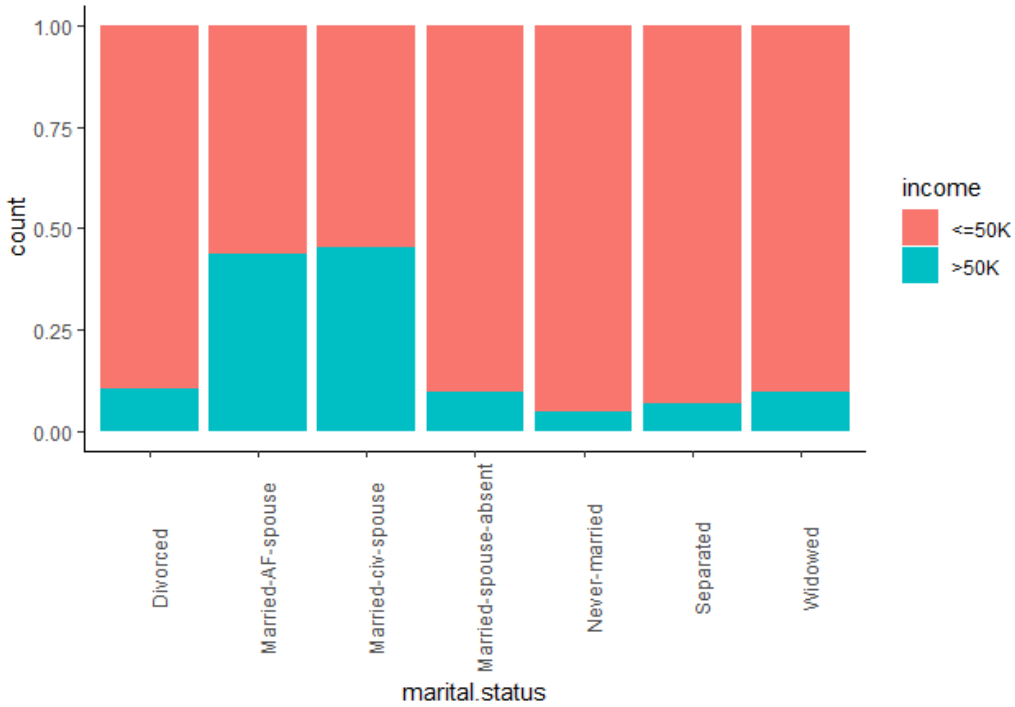


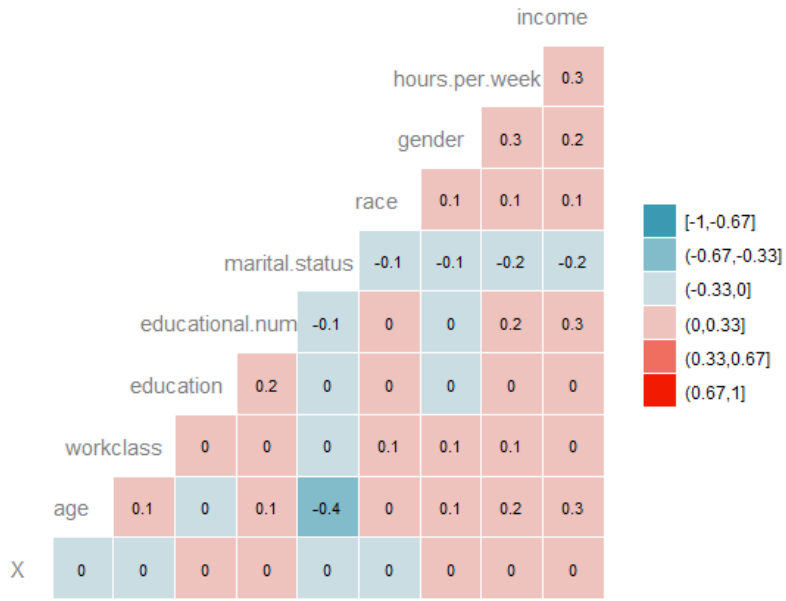


https://www.guru99.com/r-generalized-linear-model.html

R examples - Pricing GLM

Checking Data Variables e.g. factor affect earnings , variable correlation





https://www.guru99.com/r-generalized-linear-model.html

R examples - Pricing GLM

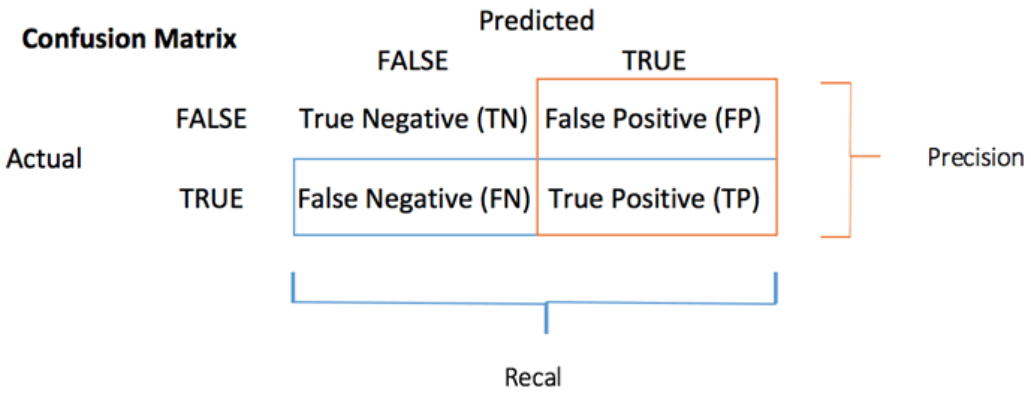
Using 80% 20% rule to split main dataset into train and test data

Using train data for model fitting

```
> formula <- income~.
> logit <- glm(formula, data = data_train, family = 'binomial')
> summary(logit)

Call:
glm(formula = formula, family = "binomial", data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6844  -0.5912  -0.2637  -0.0689   3.1938

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -6.430439   0.260757 -24.661  < 2e-16 ***
age                           0.029573   0.001392  21.238  < 2e-16 ***
workclassLocal-gov           -0.611243   0.093477  -6.539 6.19e-11 ***
workclassPrivate             -0.516368   0.078468  -6.581 4.68e-11 ***
```

https://www.guru99.com/r-generalized-linear-model.html

R examples - Pricing GLM




```
table_mat <- table(data_test$income, predict > 0.5)
> table_mat

        FALSE TRUE
  <=50K  6373  487
  >50K   1107 1240
```

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

```
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
> accuracy_Test
[1] 0.8268709
```

https://www.guru99.com/r-generalized-linear-model.html

R examples -
Data Processing and Manipulation in R

#Linking R to Excel

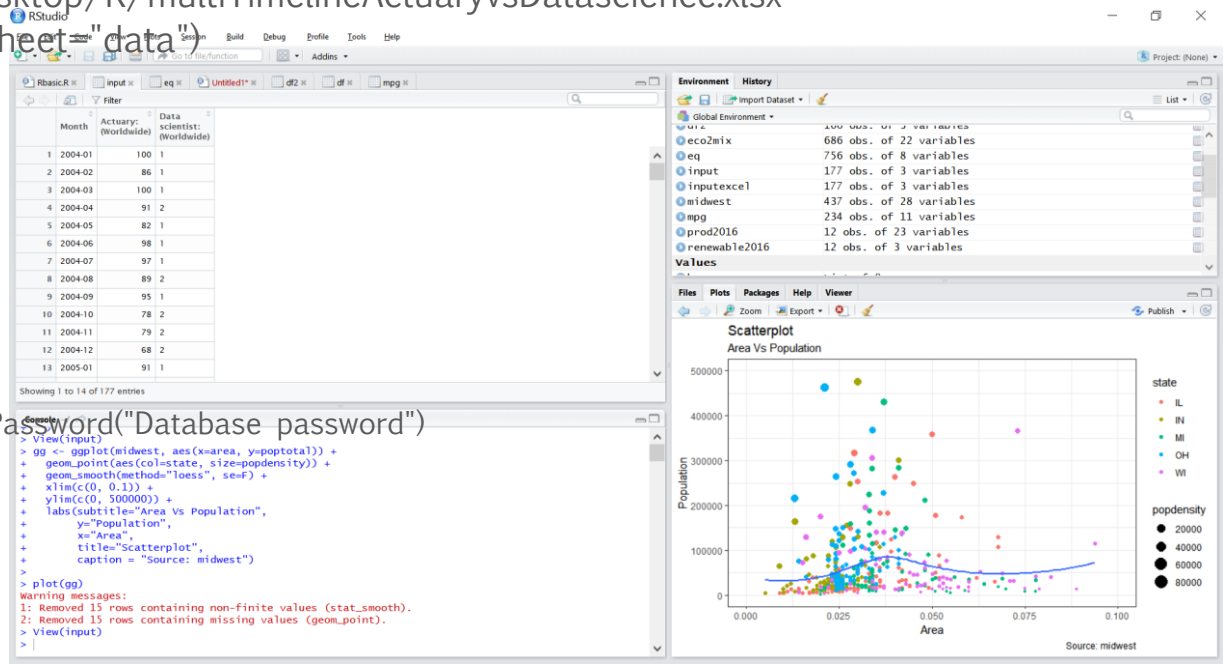
```
library(readxl)
library(readr)
examplefile<-"C:/Users/sgartiyu/Desktop/R/multiTimelineActuaryvsDatascience.csv"
input<-read_csv(examplefile)

examplefilexl<-"C:/Users/sgartiyu/Desktop/R/multiTimelineActuaryvsDatascience.xlsx"
inputexcel<-read_excel(examplefilexl,sheet="data")
```

#Linking R to SQL

```
library(odbc)
con <- dbConnect(odbc(),
                 Driver = "SQLServer",
                 Server = "mysqlhost",
                 Database = "mydbname",
                 UID = "myuser",
                 PWD = rstudioapi::askForPassword("Database password")
                 Port = 1433)
```

#Manipulation

```
#data manipulation
library(dplyr)
library(tidyr)
library(data.table)
Library(DT)
```



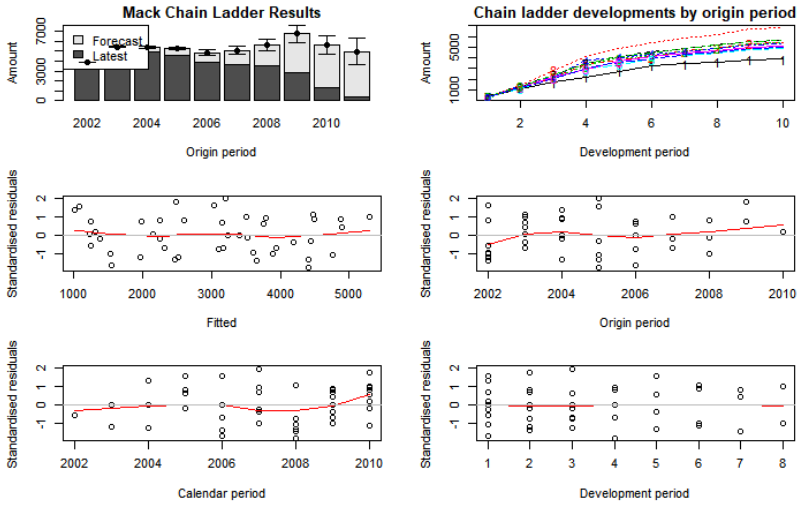
R examples - Reserving ChainLadder

- The ChainLadder package written by Markus Gesmann allows standard reserving methods to be applied to development triangles
  - Chain Ladder
  - BF Cape Cod
  - Mack Method
  - Bootstrapping
  - …

R examples - Reserving ChainLadder

dev										
origin	1	2	3	4	5	6	7	8	9	10
2002	358	1125	1735	2218	2746	3320	3466	3606	3834	3901
2003	352	1236	2170	3353	3799	4120	4648	4914	5339	
2004	291	1292	2219	3235	3986	4133	4629	4909		
2005	311	1419	2195	3757	4030	4382	4588			
2006	443	1136	2128	2898	3403	3873				
2007	396	1333	2181	2986	3692					
2008	441	1288	2420	3483						
2009	359	1421	2864							
2010	377	1363								
2011	344									

	Latest	Dev.To.Da	Ultimate	IBNR	Mack.S.E	CV(IBNR)
2002	3,901	1	3,901	0	0	NaN
2003	5,339	0.9828	5,432	93.3	71.4	0.765
2004	4,909	0.9129	5,378	468.6	118.4	0.253
2005	4,588	0.8662	5,297	708.5	130.5	0.184
2006	3,873	0.7975	4,857	983.7	260.3	0.265
2007	3,692	0.7225	5,110	1,418.10	409.9	0.289
2008	3,483	0.6154	5,659	2,176.50	557.3	0.256
2009	2,864	0.4223	6,782	3,918.00	873.9	0.223
2010	1,363	0.2417	5,640	4,277.30	970.4	0.227
2011	344	0.0693	4,967	4,623.30	1,360.90	0.294



https://gist.github.com/mages/3687713/659b2826d429823ff4ddb139d4d1bf46fe794dac
https://rawgit.com/mages/GIRO2012/master/Using_R_in_Insurance_GIRO_2012.html

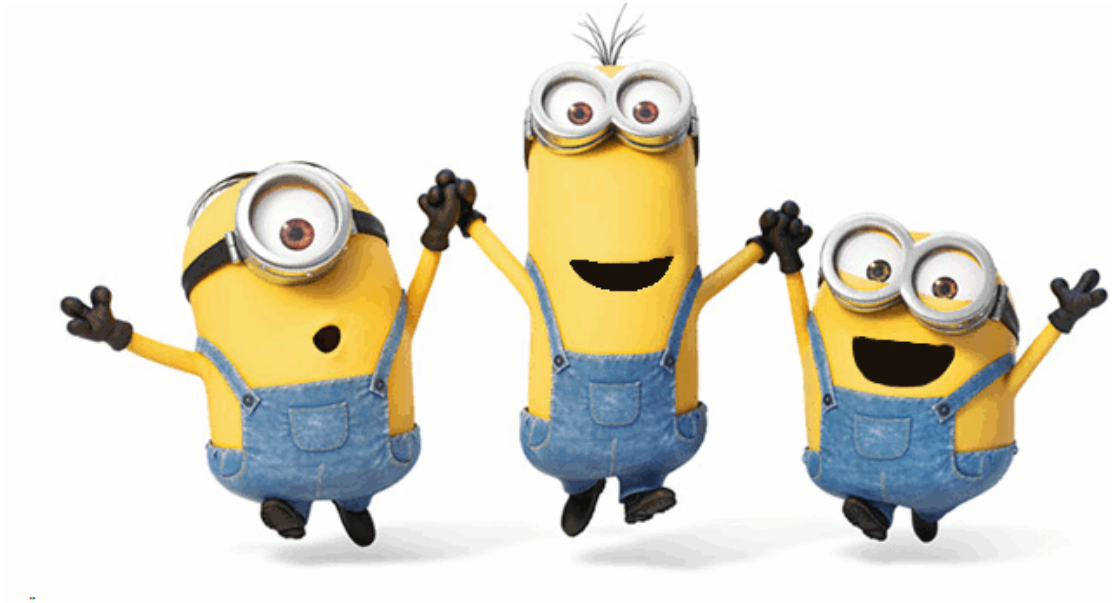
R Shiny Dashboard

- A quick, powerful way of creating and delivering management information, and a welcome user interface to R

  http://shiny.rstudio.com/

- An example : insurance frequency severity simulation tool
- https://tychobra.shinyapps.io/freq-sev-claims-sim/

Resource and next steps



Useful R packages/Tools

- *ChainLadder* for reserving triangles and methods
- *Dplyr* for data manipulation and data cleaning
- R *glm* for generalised linear modelling
- *Leaflet* for geographical analysis
- *Ggplot2* for data visualisation
- *RMarkdown* for reports
- *KNN* for spatial smoothing
- *XGBoost* for pricing and wider regression application

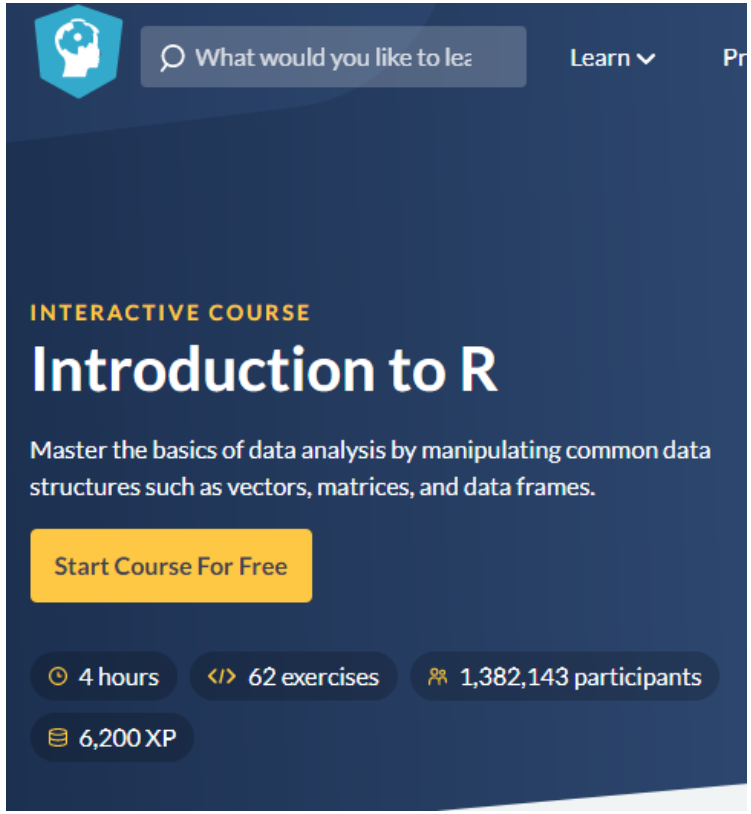
Source data and code : http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#top

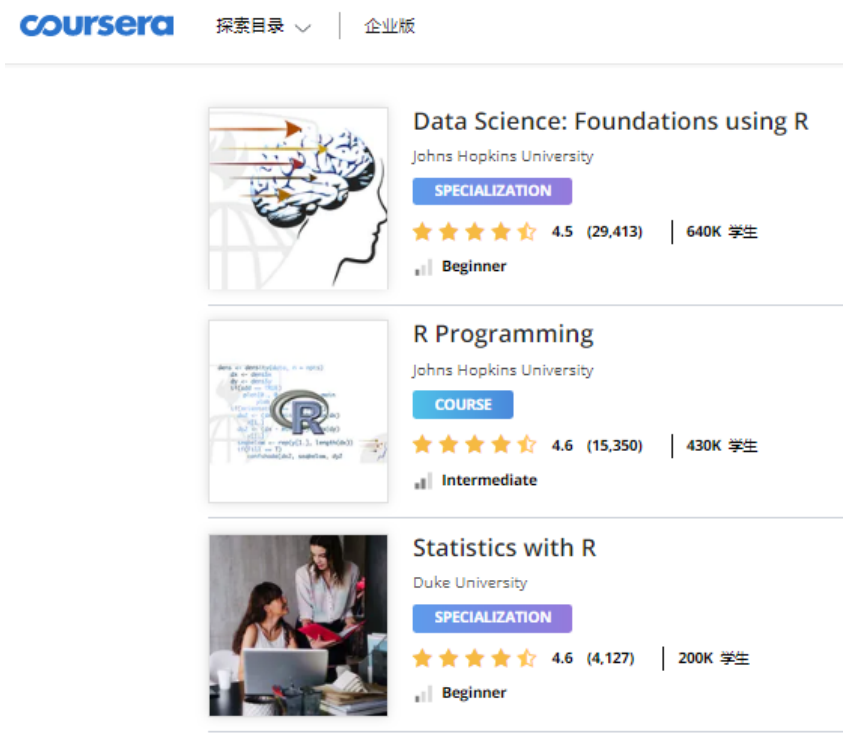
Resources

- The best way to learn a programing language is to USE IT !
- Data Science online course
- Google
- Keep practising
- Kaggle Competition

Appendix :
Online Resources

- Courses to get you started
  - DataCamp : Introduction to R (free course)
  - Coursera : Machine Learning (Andrew Ng)





Appendix :
Popular R packages for actuaries

actuar: Loss distributions modelling, risk theory (including ruin theory), simulation of compound hierarchical models and credibility theory

ChainLadder: Reserving methods in R

copula: Multivariate Dependence with Copulas

cplm: Monte Carlo EM algorithms and Bayesian methods for fitting Tweedie compound Poisson linear models

evir: Extreme Values in R

fitdistrplus: Help to fit of a parametric distribution to non-censored or censored data

lifecontingencies: Package to perform actuarial evaluation of life contingencies

lossDev: A Bayesian time series loss development model

mondate: R package to keep track of dates in terms of months