R tutorial for SS3859

Yuanhao Lai

September 10, 2018

Contents

1 What are R, RStudio, and R Markdown?

R: a free software environment for statistical computing and graphics. R was born for statistical analysis.

RStudio: a powerful, free, open-source integrated development environment for R.

R Markdown: a simple formatting syntax (Markdown) for authoring HTML, PDF, and MS Word documents with R codes embedded. R Markdown documents are fully **reproducible** and support dozens of static and dynamic output formats. It keeps evolving and it supports embeding interactive applications in a HTML page. Our goal in this course is to use it to generate a PDF document.

In particular, this document was generated by R Markdown.

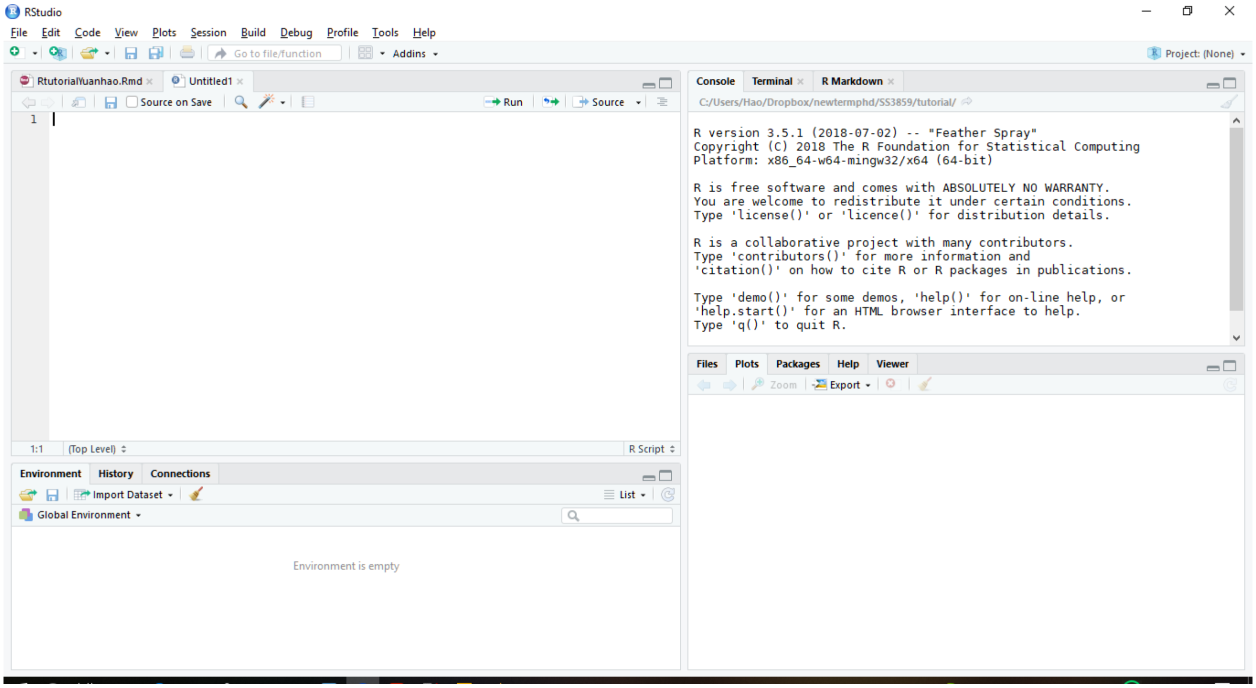
2 R and RStudio

2.1 Installation

You can download R from https://www.r-project.org and RStudio from https://www.rstudio.com/products/rstudio/download/#download. Choose the version that fits your operating system (Windows or Mac).

2.2 Use R

Once you have the R and RStudio ready, start RStudio and create a R script file.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
RtutorialYuanhao.Rmd
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Console Terminal R Markdown
C:/Users/Hao/Dropbox/newtermphd/SS3859/tutorial/
Files Plots Packages Help Viewer
Zoom
Export

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

Figure 1: Layout of RStudio

Below we show how to use R to solve **Exercise 1.3** in the textbook.

This exercise considers the **Automobile Insurance Claims data**, consisting of,

- STATE CODE: codes 01 through 17 used, with each code randomly assigned to an actual individual state
- CLASS: rating class of operator, based on age, sex, marital status, and use of vehicle
- GENDER: operator sex AGE: operator age
- PAID: amount paid to settle and close a claim.

You are focusing on older drivers, 50 and older, for which there are n = 6,773 claims available.

- Examine the histogram of the amount PAID and comment on the symmetry.
- Create a new variable, the (natural) logarithmic claims paid, LNPAID.
- Create a histogram and a qq plot of LNPAID.
- Comment on the symmetry of this variable.
- Does it appear to be approximately normally distributed? (I added this)

2.2.1 Read data

```
# Read the data from a url
df<-read.csv("http://fisher.stats.uwo.ca/faculty/aim/2018/3859A/data/AutoClaims.csv",
             header=TRUE)
```

2.2.2 View data

These are common ways to check the data at the first stage. There is a way to present the result more formally in R Markdown.

```r
# View data
str(df)
```

```
## 'data.frame':    6773 obs. of  5 variables:
##  $ STATE : Factor w/ 13 levels "STATE 01","STATE 02",..: 11 12 12 12 12 12 12 7 11 3 ...
##  $ CLASS : Factor w/ 18 levels "C1 ","C11","C1A",..: 7 7 2 16 16 16 2 7 2 2 ...
##  $ GENDER: Factor w/ 2 levels "F","M": 2 2 2 1 2 2 2 2 2 2 ...
##  $ AGE   : int  97 96 95 95 95 95 94 94 93 93 ...
##  $ PAID  : num  1134 3761 7842 2385 650 ...
```

```r
head(df,5)
```

```
##      STATE CLASS GENDER AGE    PAID
## 1 STATE 14   C6       M  97 1134.44
## 2 STATE 15   C6       M  96 3761.24
## 3 STATE 15   C11      M  95 7842.31
## 4 STATE 15   F6       F  95 2384.67
## 5 STATE 15   F6       M  95  650.00
```

2.2.3 PAID

```r
# Histogram of PAID
hist(df$PAID,xlab="PAID",main="",freq = FALSE) #freq=FALSE makes the area=1
```

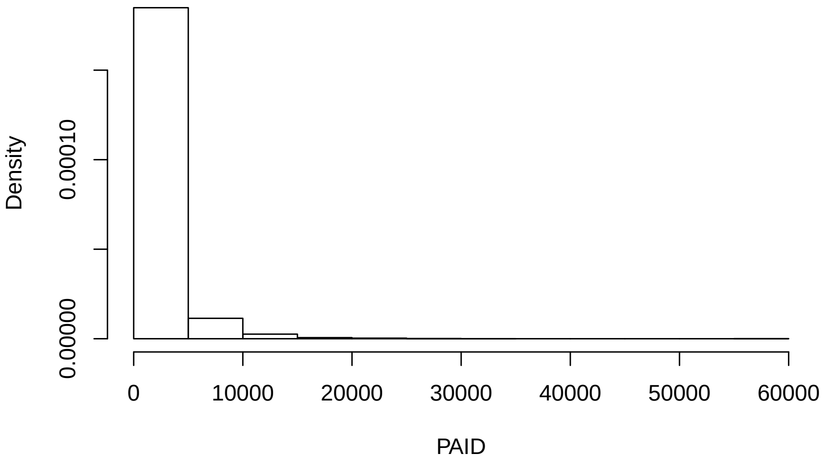


Figure 2: Histogram of PAID

Comment: The histogram of **PAID** appears to be skewed to the right. We may also call this positive skew, right-skewed, or right-tailed.

2.2.4 LNPAID

```
# The (natural) logarithmic claims paid
LNPAID <- log(df$PAID)

# Histogram of LNPAID
hist(LNPAID,xlab="LNPAID",main="",freq = FALSE)

# Add an estimtaed normal curve
curve(dnorm(x,mean=mean(LNPAID), sd=sd(LNPAID)),col="red",add = TRUE)
```

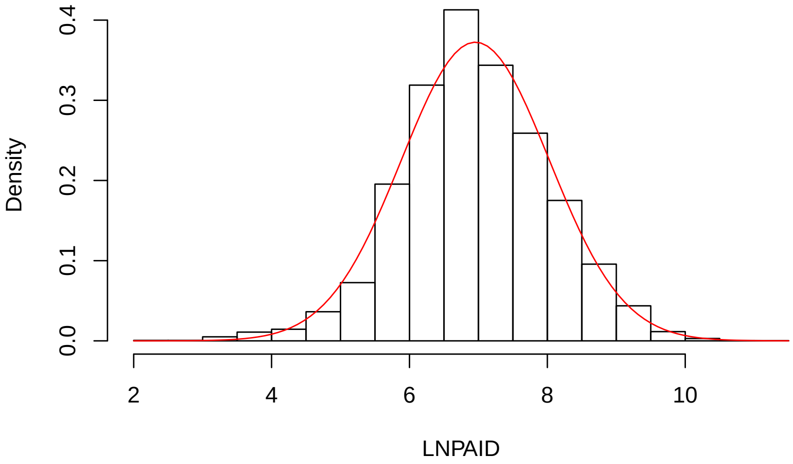


Figure 3: Histogram of LNPAID with a normal curve superimposed

```
# qqplot of LNPAID
qqnorm(LNPAID, main = "")
qqline(LNPAID)
```

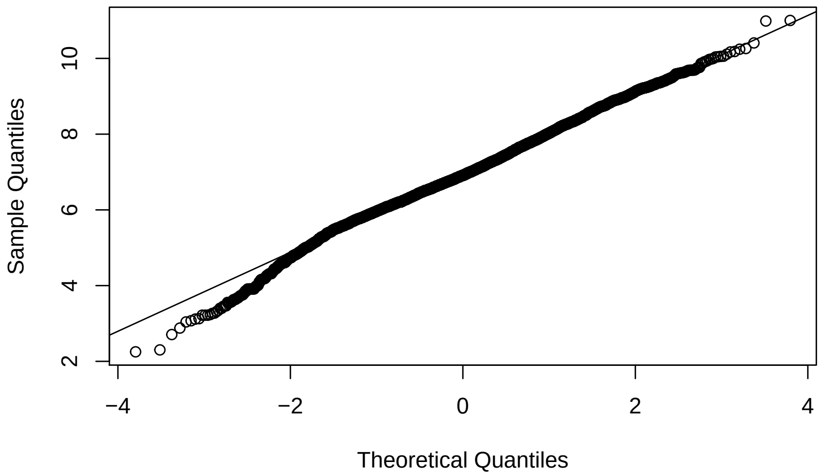


Figure 4: Normal Q-Q plot of LNPAID

Comment: Both the histogram and the QQ-plot suggest that the logarithmic transformed variable, **LNPAID**, is symmetric and close to a normal distribution.

2.3 How to learn R/data science?

Nowadays, there are a lot of resources available on the internet.

- One comprehensive R book for beginner is from Rmetrics.
- Ask and learn from *Google*? Most questions you will meet probably have appeared in *Stack Overflow*.
- If you are familiar with R and are enthusiatic in applying R on data analysis besides linear regression, I highly recommend you to take a look at Kaggle's kernel playgroud. There you can find people use R/RMarkdown or Python to create excellent documents of data analysis. You would benefit from their experiences.
- Data source for graduate students. Recently, Google announced their dataset search engine.(https://toolbox.google.com/datasetsearch)

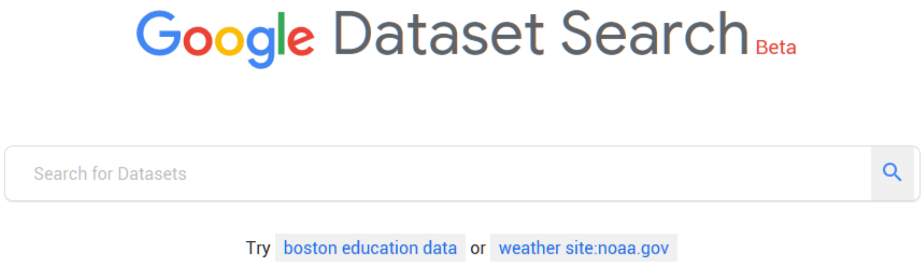

Google Dataset Search Beta
Search for Datasets
Try boston education data or weather site:noaa.gov

Figure 5: Google dataset search

3 R Markdown

3.1 Installation

A lot of people got in trouble in this step (lol). To ensure that the RStudio can produce a PDF file correctly. You need to make sure the following have been done. Besides, it is good to keep R, RStudio and related R packages up to date.

Make sure that you had R and Rstudio, then open Rstudio and,

```
# Install from CRAN
install.packages('rmarkdown')
```

In order to generate a PDF, a Latex distribution must be install. For example, MikTex for windows and MacTex for Mac. For R Markdown users who have not installed LaTeX before, the author of *rmarkdown* recommended the TinyTeX (https://yihui.name/tinytex/):

```
install.packages("tinytex")
tinytex::install_tinytex()  # install TinyTeX
```

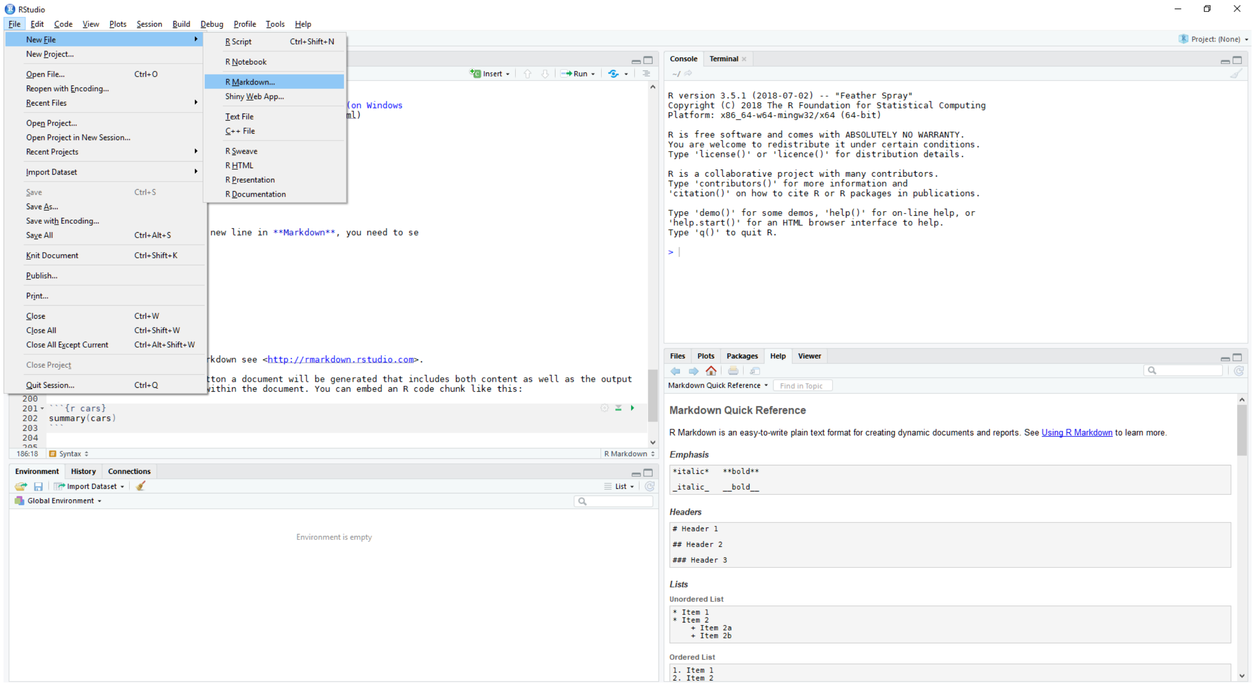
More references can be found below,

R Markdown: The Definitive Guide

Create PDF reports using R, R Markdown, LaTeX and knitr (on macOS High Sierra)

3.2 Start

We can start from creating a simple template.



```
(on Windows
ml)
```

```
new line in **Markdown**, you need to se
```

```
rkdown see <http://rmarkdown.rstudio.com>.
```

~~~
tton a document will be generated that includes both content as well as the output
within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
~~~

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Markdown Quick Reference

R Markdown is an easy-to-write plain text format for creating dynamic documents and reports. See Using R Markdown to learn more.

Emphasis

```
*italic*   **bold**

_italic_   __bold__
```

Headers

```
# Header 1

## Header 2

### Header 3
```

Lists

Unordered List

```
* Item 1
* Item 2
    + Item 2a
    + Item 2b
```

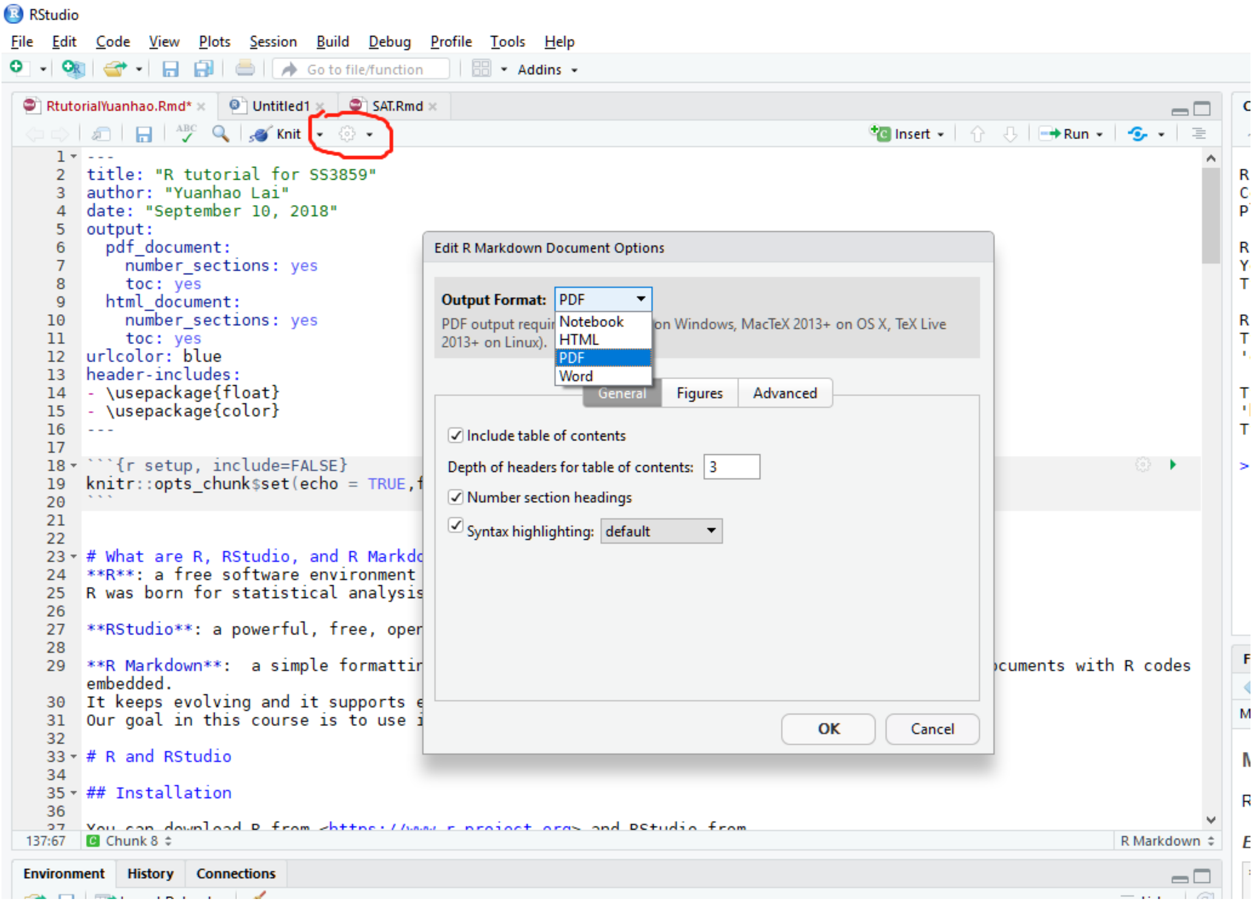
Ordered List

```
1. Item 1
2. Item 2
```

Figure 6: R Markdown template

3.3 Format

You can switch the document from a PDF to a HTML easily.



~~~
---
title: "R tutorial for SS3859"
author: "Yuanhao Lai"
date: "September 10, 2018"
output:
  pdf_document:
    number_sections: yes
    toc: yes
  html_document:
    number_sections: yes
    toc: yes
urlcolor: blue
header-includes:
- \usepackage{float}
- \usepackage{color}
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE,f
~~~

What are R, RStudio, and R Markdc
R: a free software environment
R was born for statistical analysis

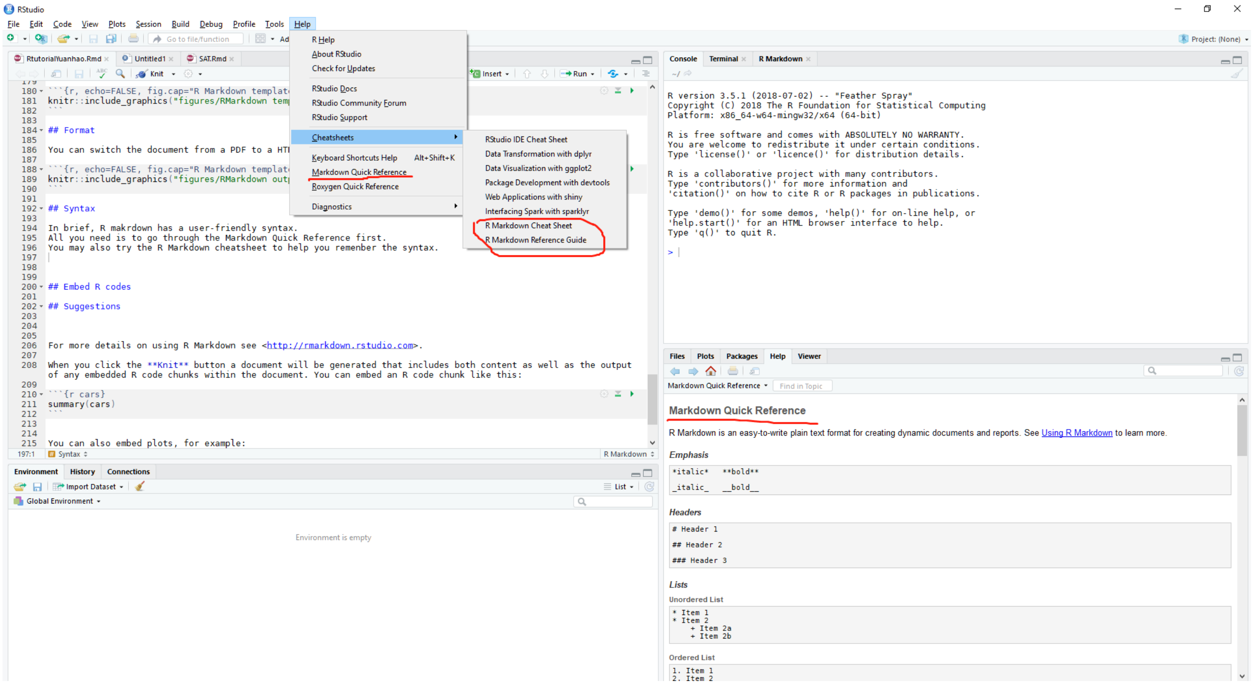
RStudio: a powerful, free, oper

R Markdown: a simple formattir   ocuments with R codes
embedded.
It keeps evolving and it supports e
Our goal in this course is to use i

R and RStudio

Installation

You can download R from <https://www.r-project.org> and RStudio from
```

Edit R Markdown Document Options

Output Format: PDF

PDF output requi… on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

☑ Include table of contents

Depth of headers for table of contents: 3

☑ Number section headings

☑ Syntax highlighting: default

Figure 7: R Markdown output format

3.4 Syntax

In brief, R makrdown has a user-friendly syntax. All you need is to go through the Markdown Quick Reference first. You may also try the R Markdown cheatsheet to help you remenber the syntax.

Try a few example and get your hands dirty in order to learn.
```



~~~
```{r, echo=FALSE, fig.cap="R Markdown templat
knitr::include_graphics("figures/RMarkdown tem
~~~

## Format

You can switch the document from a PDF to a HT

```{r, echo=FALSE, fig.cap="R Markdown templat
knitr::include_graphics("figures/RMarkdown out
```

Syntax

In brief, R makrdown has a user-friendly syntax.
All you need is to go through the Markdown Quick Reference first.
You may also try the R Markdown cheatsheet to help you remenber the syntax.

Embed R codes

Suggestions

For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

You can also embed plots, for example:
```

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

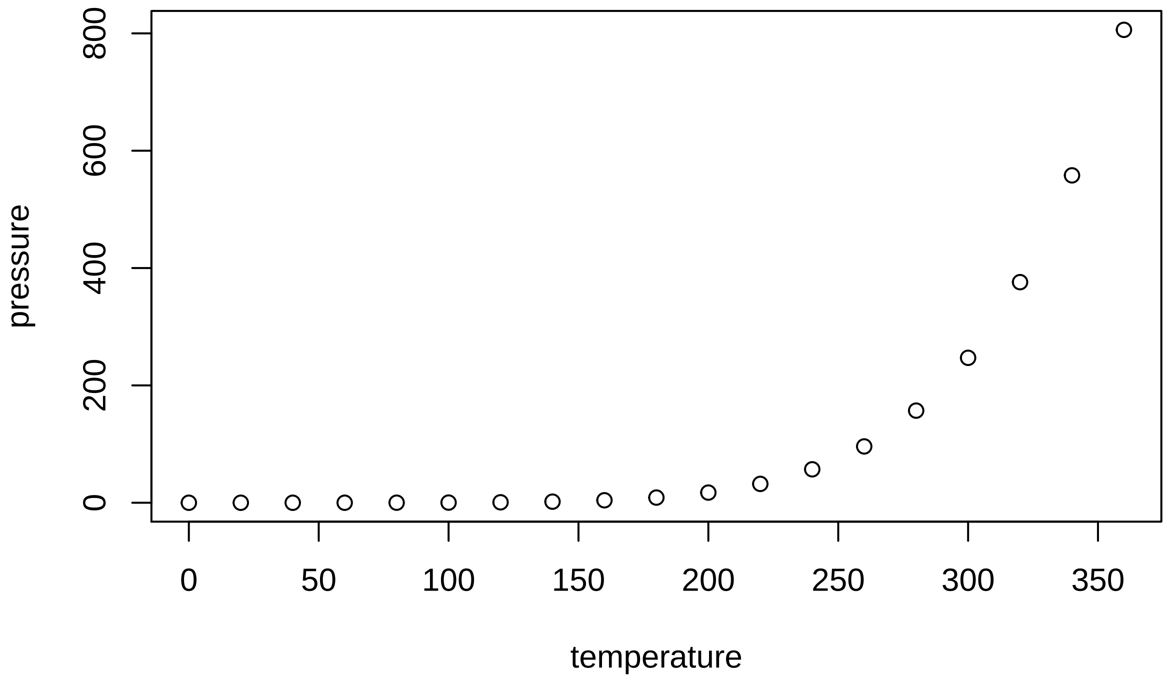
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Markdown Quick Reference

R Markdown is an easy-to-write plain text format for creating dynamic documents and reports. See Using R Markdown to learn more.

Emphasis

```
*italic*   **bold**

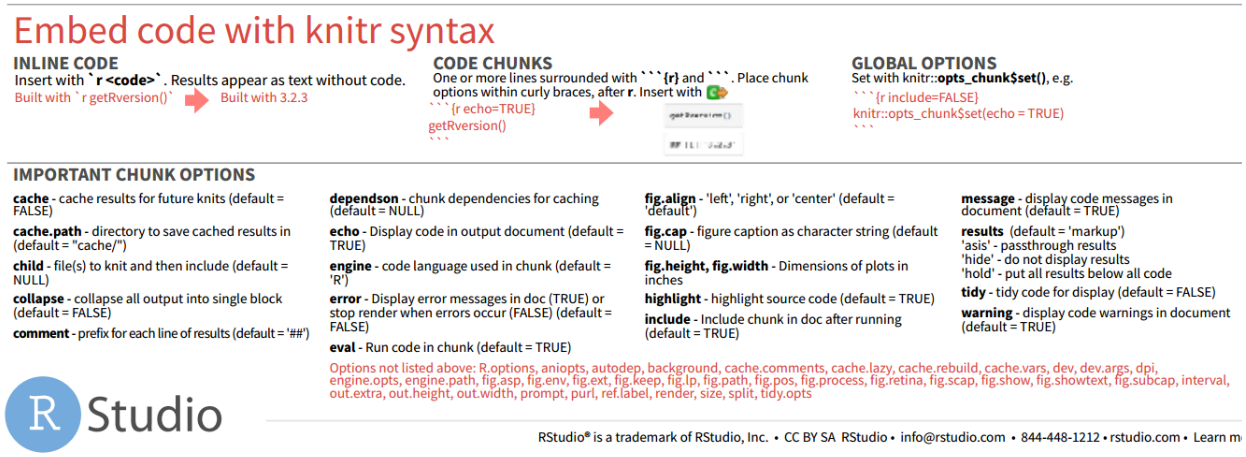
_italic_   __bold__
```

Headers

```
# Header 1

## Header 2

### Header 3
```

Lists

Unordered List

```
* Item 1
* Item 2
    + Item 2a
    + Item 2b
```

Ordered List

```
1. Item 1
2. Item 2
```

Figure 8: R Markdown help mannual

Line break:

One thing you need to pay attention is that in order to begin a new line, you need add a empty line to seperate sentences.

R Markdown will treat multiple empty lines as one single line break.

3.5 Embed R codes

There are two ways of embeding R codes in the R Markdown document, inline R code or R code Blocks.

Inline text:

There were 50 cars studied.

Code blocks,

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

You can also embed plots and hide the codes by setting the option `echo = FALSE`, for example:
```



You can find summary of the important options from the R Markdown cheatsheet.



Embed code with knitr syntax

INLINE CODE

Insert with **`r <code>`**. Results appear as text without code.

Built with `r getRversion()` → Built with 3.2.3

CODE CHUNKS

One or more lines surrounded with **```{r}** and **```**. Place chunk options within curly braces, after **r**. Insert with

~~~
```{r echo=TRUE}
getRversion()
```
~~~

GLOBAL OPTIONS

Set with knitr::**opts_chunk$set()**, e.g.

~~~
```{r include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
~~~

IMPORTANT CHUNK OPTIONS

cache - cache results for future knits (default = FALSE)

cache.path - directory to save cached results in (default = "cache/")

child - file(s) to knit and then include (default = NULL)

collapse - collapse all output into single block (default = FALSE)

comment - prefix for each line of results (default = '##')

dependson - chunk dependencies for caching (default = NULL)

echo - Display code in output document (default = TRUE)

engine - code language used in chunk (default = 'R')

error - Display error messages in doc (TRUE) or stop render when errors occur (FALSE) (default = FALSE)

eval - Run code in chunk (default = TRUE)

fig.align - 'left', 'right', or 'center' (default = 'default')

fig.cap - figure caption as character string (default = NULL)

fig.height, fig.width - Dimensions of plots in inches

highlight - highlight source code (default = TRUE)

include - Include chunk in doc after running (default = TRUE)

message - display code messages in document (default = TRUE)

results (default = 'markup')
'asis' - passthrough results
'hide' - do not display results
'hold' - put all results below all code

tidy - tidy code for display (default = FALSE)

warning - display code warnings in document (default = TRUE)

Options not listed above: R.options, aniopts, autodep, background, cache.comments, cache.lazy, cache.rebuild, cache.vars, dev, dev.args, dpi, engine.opts, engine.path, fig.asp, fig.env, fig.ext, fig.keep, fig.lp, fig.path, fig.pos, fig.process, fig.retina, fig.scap, fig.show, fig.showtext, fig.subcap, interval, out.extra, out.height, out.width, prompt, purl, ref.label, render, size, split, tidy.opts

RStudio® is a trademark of RStudio, Inc. • CC BY SA RStudio • info@rstudio.com • 844-448-1212 • rstudio.com • Learn m

Figure 9: R Markdown chunk option

3.6 Tables

Markdown has its own syntax to create a table, you may use,

First Header	Second Header
Content Cell	Content Cell
Content Cell	Content Cell

You may use a Markdown Tables Generator to do this.

To generate a table for the PDF/HTML from a R output, you may use the `stargazer` package,

```r
library(stargazer)
df<-read.csv("http://fisher.stats.uwo.ca/faculty/aim/2018/3859A/data/AutoClaims.csv",
             header=TRUE)

# For pdf
stargazer(head(df,5), type="latex", title="First 5 observations", header=TRUE, summary=FALSE )
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Sep 11, 2018 - 4:18:52 PM

Table 2: First 5 observations

	STATE	CLASS	GENDER	AGE	PAID
1	STATE 14	C6	M	97	1,134.440
2	STATE 15	C6	M	96	3,761.240
3	STATE 15	C11	M	95	7,842.310
4	STATE 15	F6	F	95	2,384.670
5	STATE 15	F6	M	95	650

```r
# For html
#stargazer(head(df,5), type="html", title="First 5 observations", header=TRUE, summary=FALSE )
```

There are more options of achieving this such as using the `kable()` fucntion.

3.7 Suggestions

- The book, R Markdown: The Definitive Guide provides a comprehensive usage of R Markdown. You may also use R Markdown to generate an interactive presentation slide or write a paper.
- RStudio provides a useful tutorial.
- As your TA, I provide office hours by an email appointment (ylai72@uwo.ca, WSC236). I will often be available on Thursday.
- Suggestions are welcome if you want me to explain something particular during the tutorial.