Introduction To R

Chris Chia

November 24, 2019

What is R? Why R?

R is a programming language that allows you to perform statistics and visualisations. R is excellent because:

- Used widely among statisticians and mathematicians and gaining traction amongst economists, political scientists, social scientists
- Large community means - can easily troubleshoot any problems online
- Large collection of in-built libraries/tools to support your data analysis needs - beautiful data visualisation tools.
- More flexible, versatile than Stata

Outline

We're going to try and achieve these two outcomes in this 20 minute tutorial:

1. Introduction: What is R is and why use R?

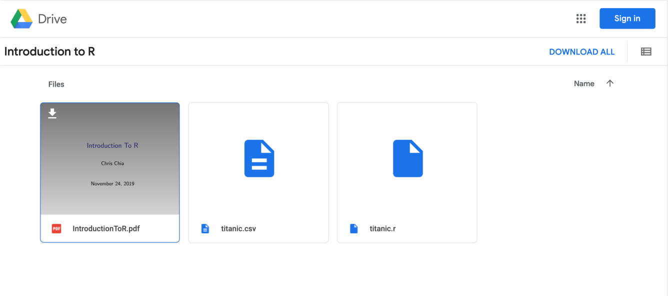
2. Explore data storytelling by getting hands on with R and create data visualisations

Let’s get hands-on

- We want to ask questions, and use data to answer them - but also in a visual and interpretable way
- We’ll use the ”Titanic” dataset (from Stanford CS109), a table of passengers and their information (sex, age, family, passenger class) - and whether they survived.
- We’ll explore whether **survival rate** is different for different **passenger classes**

R and Data Setup

- Open RStudio. If you don't have it installed, download it at `http://rstudio.com`
- Download the titanic dataset (under the file name "titanic.csv"), and the file "titanic.R" from this url: `http://shorturl.at/fkSVW`



Introduction to R

Introduction To R
Chris Chia
November 24, 2019

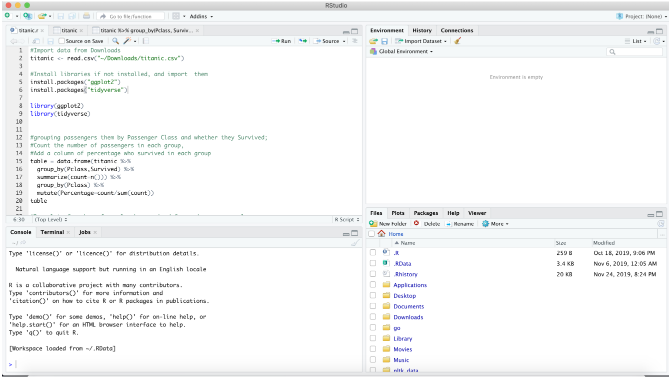
IntroductionToR.pdf

titanic.csv

titanic.r

Setup

Open the "titanic.r" file in RStudio, which will contain the commands we want to use to create data visualisations.
from the menu File → Open File. (or Cmd+O on Mac, Ctrl + O on Windows)
You should see a screen like this:



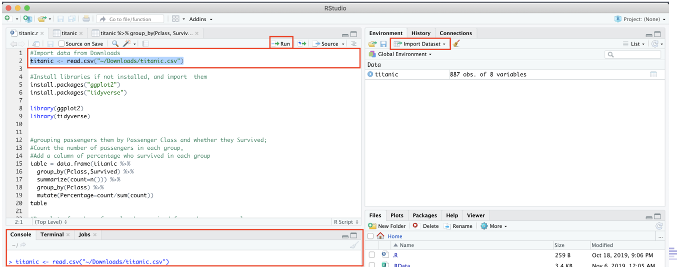
```r
#Import data from Downloads
titanic <- read.csv("~/Downloads/titanic.csv")

#Install libraries if not installed, and import  them
install.packages("ggplot2")
install.packages("tidyverse")

library(ggplot2)
library(tidyverse)


#grouping passengers them by Passenger Class and whether they Survived;
#Count the number of passengers in each group,
#Add a column of percentage who survived in each group
table = data.frame(titanic %>%
  group_by(Pclass,Survived) %>%
  summarize(count=n())) %>%
  group_by(Pclass) %>%
  mutate(Percentage=count/sum(count))
table
```

```
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

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.

[Workspace loaded from ~/.RData]

> 
```

Name	Size	Modified
.R	259 B	Oct 18, 2019, 9:06 PM
.RData	3.4 KB	Nov 6, 2019, 12:05 AM
.Rhistory	20 KB	Nov 24, 2019, 8:24 PM
Applications		
Desktop		
Documents		
Downloads		
go		
Library		
Movies		
Music		
nltk_data		

Data Import

- In the code editor, select the line ”**titanic ← read.csv(”∼/Downloads/titanic.csv”)**” . Make sure to change **∼/Downloads/** to the folder where you saved ”titanic.csv”, and then press the ”Run” button. This assigns the dataset to a variable (name) ”titanic” by executing the command on the console.
- If that does not work, try importing it directly from the environment pane menu Import Dataset → From Text(base)
- This is why we use a ”.r” file - so that others can reuse and replicate the steps we took. Also, all the commands need to be run again when we restart



```
#Import data from Downloads
titanic <- read.csv("~/Downloads/titanic.csv")

#Install libraries if not installed, and import  them
install.packages("ggplot2")
install.packages("tidyverse")

library(ggplot2)
library(tidyverse)


#grouping passengers them by Passenger Class and whether they Survived;
#Count the number of passengers in each group,
#Add a column of percentage who survived in each group
table = data.frame(titanic %>%
  group_by(Pclass,Survived) %>%
  summarize(count=n())) %>%
  group_by(Pclass) %>%
  mutate(Percentage=count/sum(count))
table

```

```
> titanic <- read.csv("~/Downloads/titanic.csv")
```

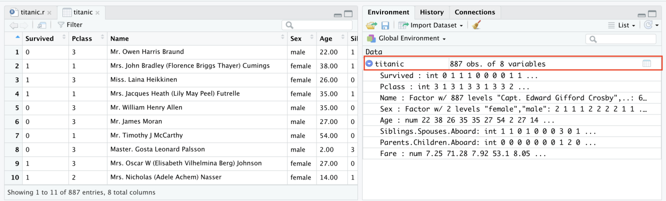
Data

titanic	887 obs. of 8 variables

Name	Size	Modified
.R	259 B	Oct 18, 2019, 9:06 PM
.RData	3.4 KB	Nov 6, 2019, 12:05 AM

Exploring the Data

In the environment pane, press the right arrow icon, and click on the titanic rectangle. This is the dataset in raw table form. You should see something like this:



	Survived	Pclass	Name	Sex	Age	Sil
1	0	3	Mr. Owen Harris Braund	male	22.00	1
2	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings	female	38.00	1
3	1	3	Miss. Laina Heikkinen	female	26.00	0
4	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.00	1
5	0	3	Mr. William Henry Allen	male	35.00	0
6	0	3	Mr. James Moran	male	27.00	0
7	0	1	Mr. Timothy J McCarthy	male	54.00	0
8	0	3	Master. Gosta Leonard Palsson	male	2.00	3
9	1	3	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	female	27.00	0
10	1	2	Mrs. Nicholas (Adele Achem) Nasser	female	14.00	1

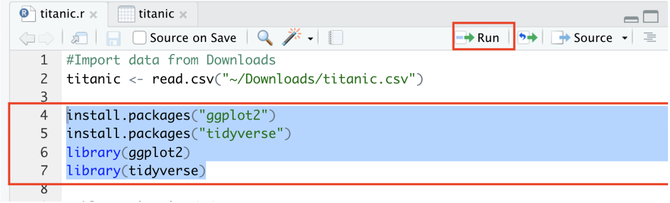
Showing 1 to 11 of 887 entries, 8 total columns

```
Data
titanic              887 obs. of 8 variables
    Survived : int 0 1 1 1 0 0 0 0 1 1 ...
    Pclass : int 3 1 3 1 3 3 1 3 3 2 ...
    Name : Factor w/ 887 levels "Capt. Edward Gifford Crosby",..: 6…
    Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 .…
    Age : num 22 38 26 35 35 27 54 2 27 14 ...
    Siblings.Spouses.Aboard: int 1 1 0 1 0 0 0 3 0 1 ...
    Parents.Children.Aboard: int 0 0 0 0 0 0 0 1 2 0 ...
    Fare : num 7.25 71.28 7.92 53.1 8.05 ...
```

We can see that it has information about passengers, their Passenger Class (**PClass**) which is ordered from 1-3 (highest to lowest), **Sex** (male, female 1 or 2), **Age**, Family information, and whether they survived.

Import libraries

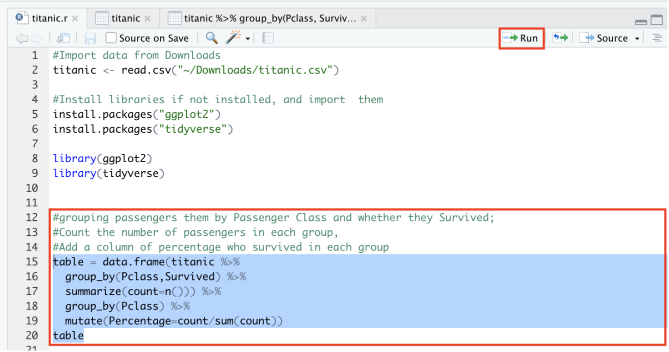
We need to install and import libraries for the data visualisation. Select these lines in the code editor, then press 'Run'.



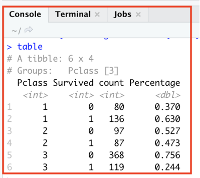
```r
#Import data from Downloads
titanic <- read.csv("~/Downloads/titanic.csv")

install.packages("ggplot2")
install.packages("tidyverse")
library(ggplot2)
library(tidyverse)
```

Exploring Data

- Select the following lines, which groups the data by passenger class and survival, and counts how many are in each category, in the code editor, then press 'Run'. You should see this output in your console:



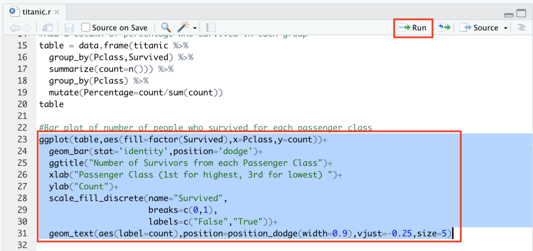
```
#Import data from Downloads
titanic <- read.csv("~/Downloads/titanic.csv")

#Install libraries if not installed, and import  them
install.packages("ggplot2")
install.packages("tidyverse")

library(ggplot2)
library(tidyverse)


#grouping passengers them by Passenger Class and whether they Survived;
#Count the number of passengers in each group,
#Add a column of percentage who survived in each group
table = data.frame(titanic %>%
  group_by(Pclass,Survived) %>%
  summarize(count=n())) %>%
  group_by(Pclass) %>%
  mutate(Percentage=count/sum(count))
table
```



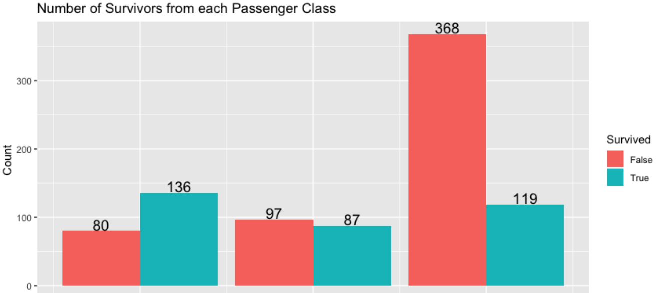
```
> table
# A tibble: 6 x 4
# Groups:   Pclass [3]
  Pclass Survived count Percentage
   <int>    <int> <int>      <dbl>
1      1        0    80      0.370
2      1        1   136      0.630
3      2        0    97      0.527
4      2        1    87      0.473
5      3        0   368      0.756
6      3        1   119      0.244
```

First Data Visualisation - Grouped Bar Plot

Now select the following lines, which creates a bar plot of the number of each people grouped by each passenger class and whether they survived

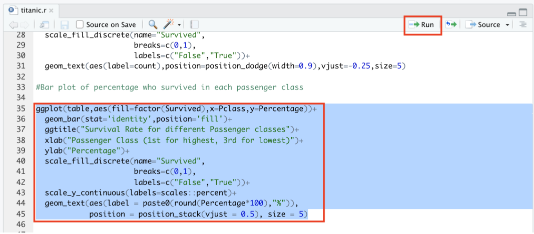


```
#Add a column of percentage who survived in each group
table = data.frame(titanic %>%
  group_by(Pclass,Survived) %>%
  summarize(count=n())) %>%
  group_by(Pclass) %>%
  mutate(Percentage=count/sum(count))
table

#Bar plot of number of people who survived for each passenger class
ggplot(table,aes(fill=factor(Survived),x=Pclass,y=count))+
  geom_bar(stat='identity',position='dodge')+
  ggtitle("Number of Survivors from each Passenger Class")+
  xlab("Passenger Class (1st for highest, 3rd for lowest) ")+
  ylab("Count")+
  scale_fill_discrete(name="Survived",
                      breaks=c(0,1),
                      labels=c("False","True"))+
  geom_text(aes(label=count),position=position_dodge(width=0.9),vjust=-0.25,size=5)
```

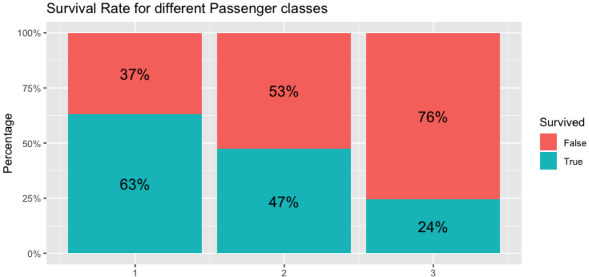


Second Data Visualisation - Stacked Percentage Bar Plot

We want to be able to make a comparison among ticket classes.Select the following lines, which will create a bar plot of survival in each passenger class as a percentage



```
  scale_fill_discrete(name="Survived",
                      breaks=c(0,1),
                      labels=c("False","True"))+
  geom_text(aes(label=count),position=position_dodge(width=0.9),vjust=-0.25,size=5)

#Bar plot of percentage who survived in each passenger class

ggplot(table,aes(fill=factor(Survived),x=Pclass,y=Percentage))+
  geom_bar(stat='identity',position='fill')+
  ggtitle("Survival Rate for different Passenger classes")+
  xlab("Passenger Class (1st for highest, 3rd for lowest)")+
  ylab("Percentage")+
  scale_fill_discrete(name="Survived",
                      breaks=c(0,1),
                      labels=c("False","True"))+
  scale_y_continuous(labels=scales::percent)+
  geom_text(aes(label = paste0(round(Percentage*100),"%")),
            position = position_stack(vjust = 0.5), size = 5)
```



Extension/Challenge

See if survival rates are different for passengers of different Sex

Further Development

- Explore relationships between survival and the other factors, e.g. Age and Family Members
- Look into building a logistic regression model with R to examine the relationship of each factor with survival
- Explore the Titanic dataset on Kaggle, a data science competition and learning platform