Introduction

STAT5002

The University of Sydney

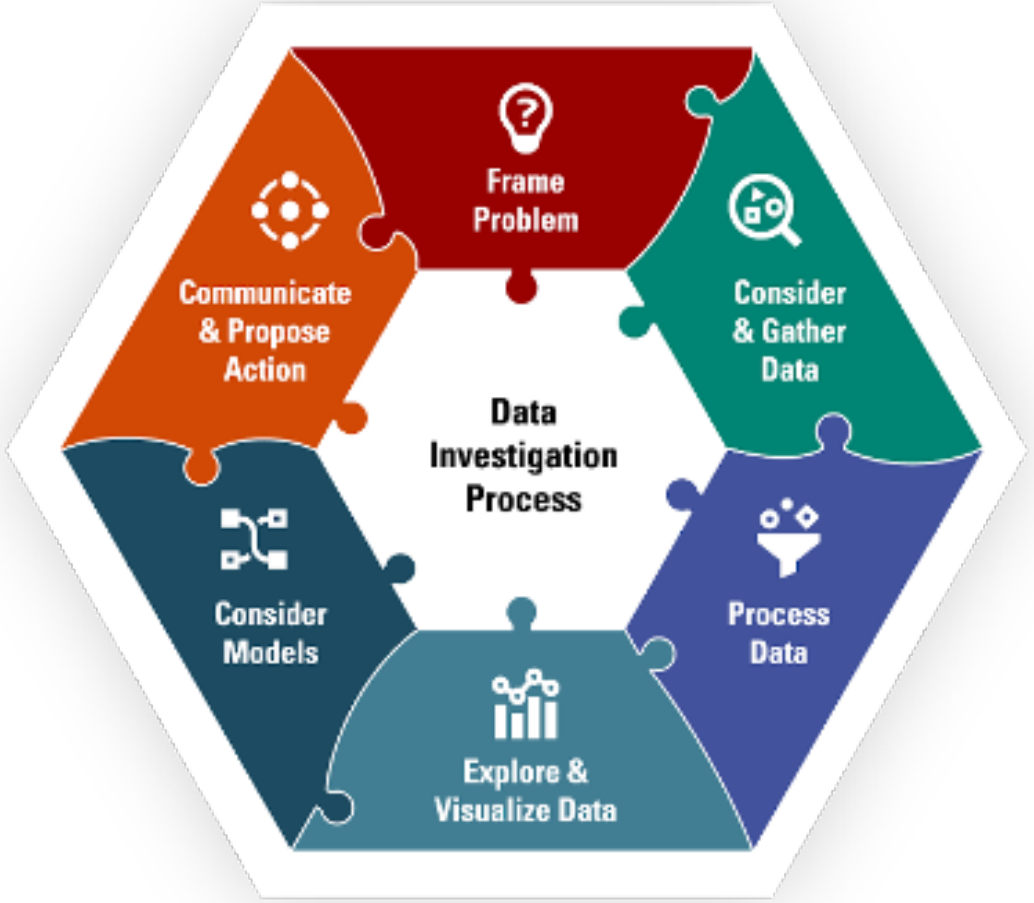
Feb 2025

THE UNIVERSITY OF SYDNEY

What is Statistics?

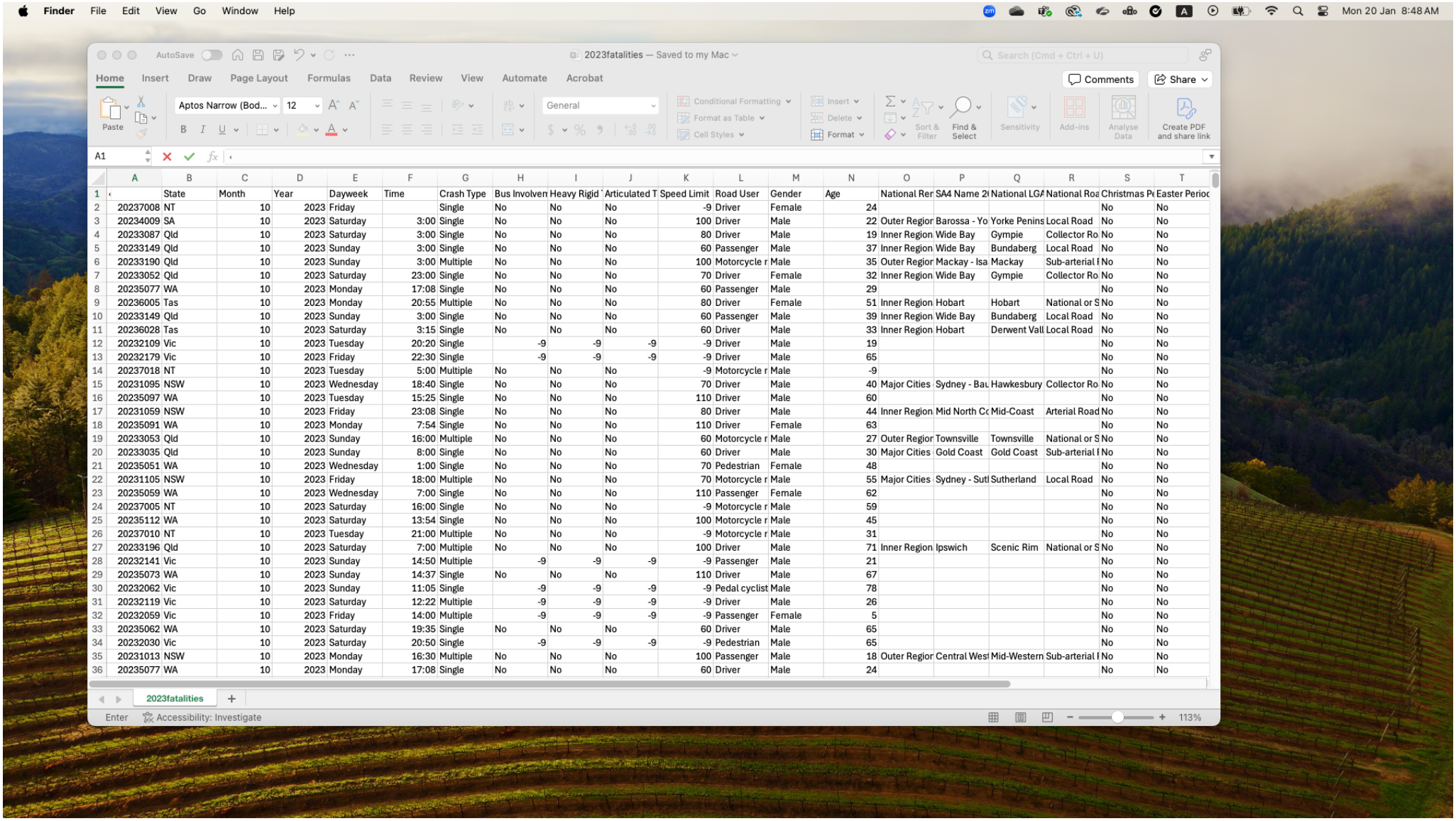
- Statistics provide powerful **quantitative tools** to **solve problems** and **make informed decisions** in a very diverse range of **real-life applications**.
- We're going to learn to **explore, visualise**, and **analyse data** to understand natural phenomena using **R** (more on this later).
- This is a unit on introduction to statistics, with an emphasis on **statistical thinking**.

Data Investigation Process



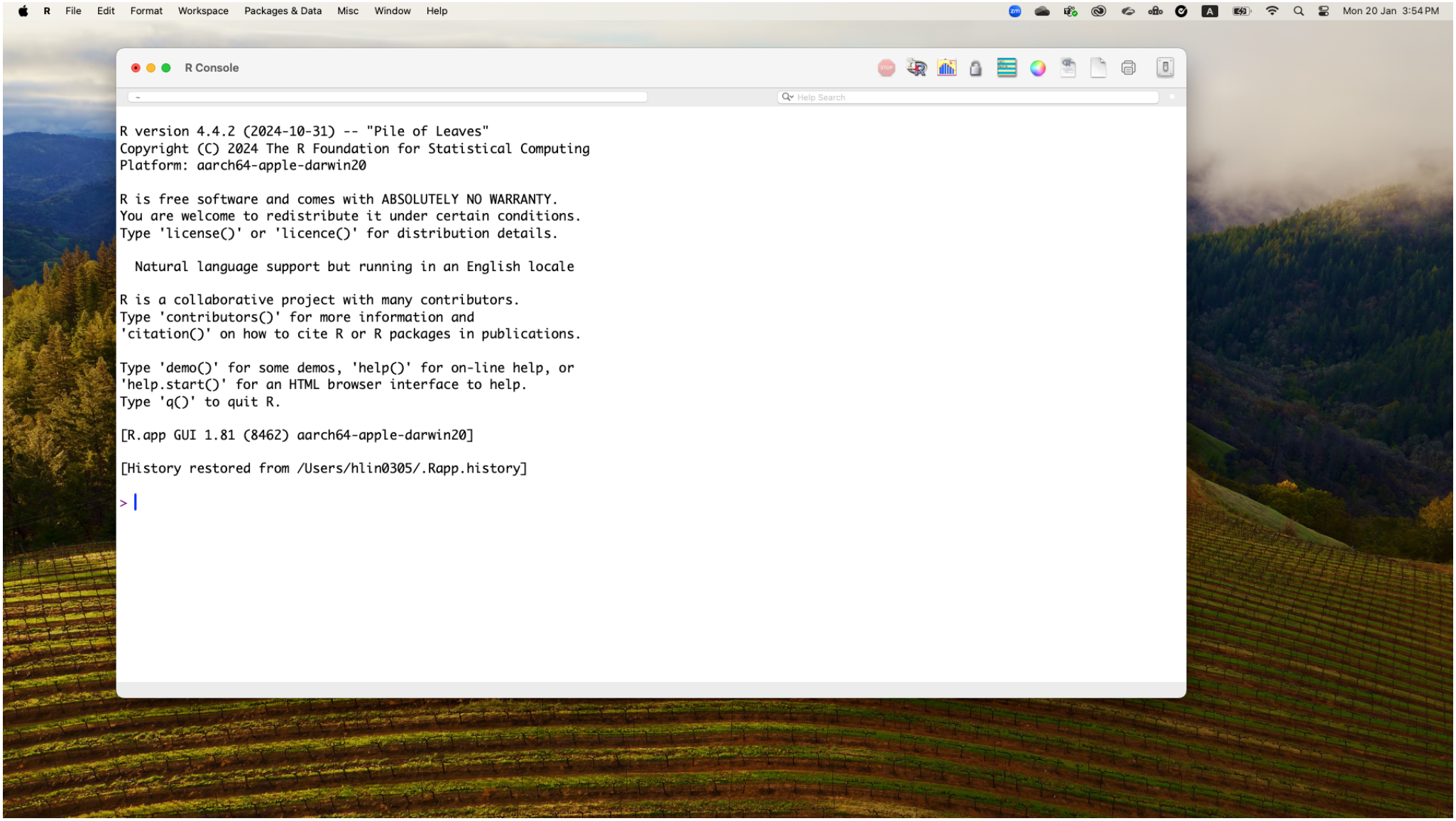
Software

Excel - not…

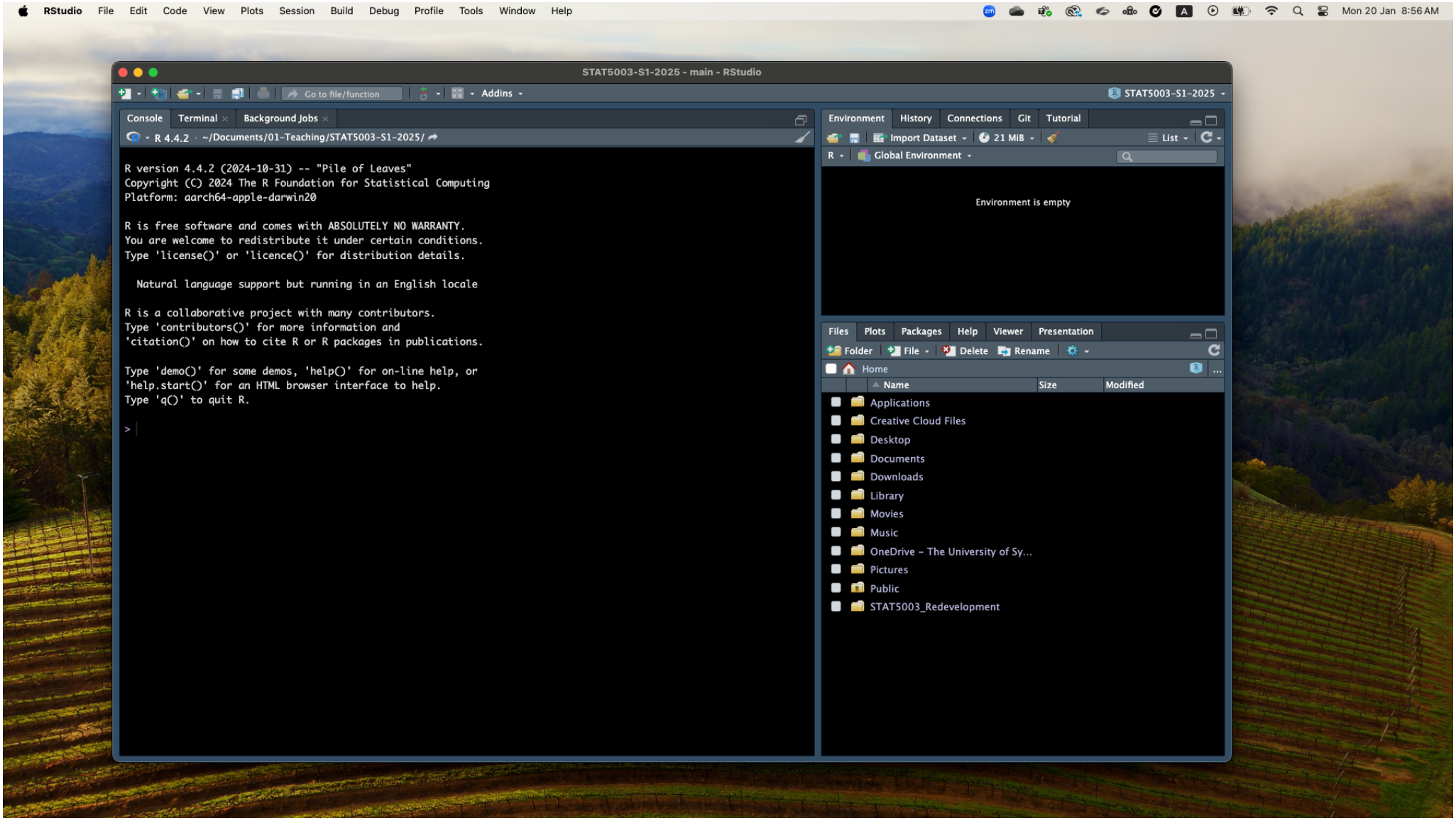


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	‹	State	Month	Year	Dayweek	Time	Crash Type	Bus Involven	Heavy Rigid '	Articulated T	Speed Limit	Road User	Gender	Age	National Rer	SA4 Name 2(	National LGA	National Roa	Christmas Pe	Easter Perioc
2	20237008	NT	10	2023	Friday		Single	No	No	No	-9	Driver	Female	24					No	No
3	20234009	SA	10	2023	Saturday	3:00	Single	No	No	No	100	Driver	Male	22	Outer Regior	Barossa - Yo	Yorke Penins	Local Road	No	No
4	20233087	Qld	10	2023	Saturday	3:00	Single	No	No	No	80	Driver	Male	19	Inner Region	Wide Bay	Gympie	Collector Ro	No	No
5	20233149	Qld	10	2023	Sunday	3:00	Single	No	No	No	60	Passenger	Male	37	Inner Region	Wide Bay	Bundaberg	Local Road	No	No
6	20233190	Qld	10	2023	Sunday	3:00	Multiple	No	No	No	100	Motorcycle r	Male	35	Outer Regior	Mackay - Isa	Mackay	Sub-arterial I	No	No
7	20233052	Qld	10	2023	Saturday	23:00	Single	No	No	No	70	Driver	Female	32	Inner Region	Wide Bay	Gympie	Collector Ro	No	No
8	20235077	WA	10	2023	Monday	17:08	Single	No	No	No	60	Passenger	Male	29					No	No
9	20236005	Tas	10	2023	Monday	20:55	Multiple	No	No	No	80	Driver	Female	51	Inner Region	Hobart	Hobart	National or S	No	No
10	20233149	Qld	10	2023	Sunday	3:00	Single	No	No	No	60	Passenger	Male	39	Inner Region	Wide Bay	Bundaberg	Local Road	No	No
11	20236028	Tas	10	2023	Saturday	3:15	Single	No	No	No	60	Driver	Male	33	Inner Region	Hobart	Derwent Vall	Local Road	No	No
12	20232109	Vic	10	2023	Tuesday	20:20	Single	-9	-9	-9	-9	Driver	Male	19					No	No
13	20232179	Vic	10	2023	Friday	22:30	Single	-9	-9	-9	-9	Driver	Male	65					No	No
14	20237018	NT	10	2023	Tuesday	5:00	Multiple	No	No	No	-9	Motorcycle r	Male	-9					No	No
15	20231095	NSW	10	2023	Wednesday	18:40	Single	No	No	No	70	Driver	Male	40	Major Cities	Sydney - Bal	Hawkesbury	Collector Ro	No	No
16	20235097	WA	10	2023	Tuesday	15:25	Single	No	No	No	110	Driver	Male	60					No	No
17	20231059	NSW	10	2023	Friday	23:08	Single	No	No	No	80	Driver	Male	44	Inner Region	Mid North Cc	Mid-Coast	Arterial Road	No	No
18	20235091	WA	10	2023	Monday	7:54	Single	No	No	No	110	Driver	Female	63					No	No
19	20233053	Qld	10	2023	Sunday	16:00	Multiple	No	No	No	60	Motorcycle r	Male	27	Outer Regior	Townsville	Townsville	National or S	No	No
20	20233035	Qld	10	2023	Sunday	8:00	Single	No	No	No	60	Driver	Male	30	Major Cities	Gold Coast	Gold Coast	Sub-arterial I	No	No
21	20235051	WA	10	2023	Wednesday	1:00	Single	No	No	No	70	Pedestrian	Female	48					No	No
22	20231105	NSW	10	2023	Friday	18:00	Multiple	No	No	No	70	Motorcycle r	Male	55	Major Cities	Sydney - Sut	Sutherland	Local Road	No	No
23	20235059	WA	10	2023	Wednesday	7:00	Single	No	No	No	110	Passenger	Female	62					No	No
24	20237005	NT	10	2023	Saturday	16:00	Single	No	No	No	-9	Motorcycle r	Male	59					No	No
25	20235112	WA	10	2023	Saturday	13:54	Single	No	No	No	100	Motorcycle r	Male	45					No	No
26	20237010	NT	10	2023	Tuesday	21:00	Multiple	No	No	No	-9	Motorcycle r	Male	31					No	No
27	20233196	Qld	10	2023	Saturday	7:00	Multiple	No	No	No	100	Driver	Male	71	Inner Region	Ipswich	Scenic Rim	National or S	No	No
28	20232141	Vic	10	2023	Sunday	14:50	Multiple	-9	-9	-9	-9	Passenger	Male	21					No	No
29	20235073	WA	10	2023	Sunday	14:37	Single	No	No	No	110	Driver	Male	67					No	No
30	20232062	Vic	10	2023	Sunday	11:05	Single	-9	-9	-9	-9	Pedal cyclist	Male	78					No	No
31	20232119	Vic	10	2023	Saturday	12:22	Multiple	-9	-9	-9	-9	Driver	Male	26					No	No
32	20232059	Vic	10	2023	Friday	14:00	Multiple	-9	-9	-9	-9	Passenger	Female	5					No	No
33	20235062	WA	10	2023	Saturday	19:35	Single	No	No	No	60	Driver	Male	65					No	No
34	20232030	Vic	10	2023	Saturday	20:50	Single	-9	-9	-9	-9	Pedestrian	Male	65					No	No
35	20231013	NSW	10	2023	Monday	16:30	Multiple	No	No	No	100	Passenger	Male	18	Outer Regior	Central West	Mid-Western	Sub-arterial I	No	No
36	20235077	WA	10	2023	Monday	17:08	Single	No	No	No	60	Driver	Male	24					No	No

R


R File Edit Format Workspace Packages & Data Misc Window Help
Mon 20 Jan 3:54 PM
R Console
Help Search
R version 4.4.2 (2024-10-31) -- "Pile of Leaves"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
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.
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.
[R.app GUI 1.81 (8462) aarch64-apple-darwin20]
[History restored from /Users/hlin0305/.Rapp.history]
>

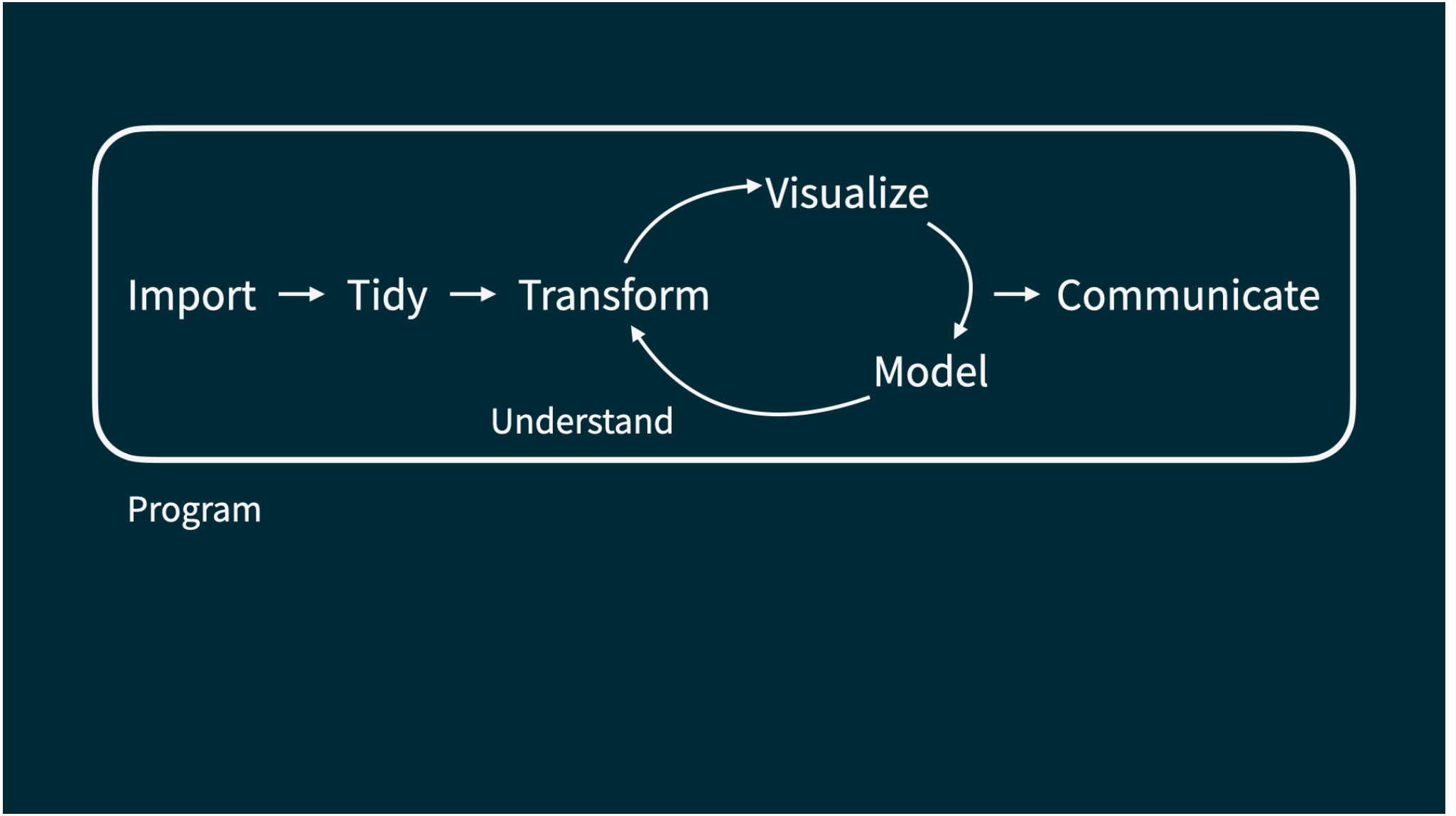
RStudio


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Mon 20 Jan 8:56 AM
STAT5003-S1-2025 - main - RStudio
Go to file/function
Addins
STAT5003-S1-2025
Console
Terminal
Background Jobs
R 4.4.2 · ~/Documents/01-Teaching/STAT5003-S1-2025/
R version 4.4.2 (2024-10-31) -- "Pile of Leaves"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
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.
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.
>
Environment
History
Connections
Git
Tutorial
Import Dataset
21 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
Folder
File
Delete
Rename
Home
Name
Size
Modified
Applications
Creative Cloud Files
Desktop
Documents
Downloads
Library
Movies
Music
OneDrive - The University of Sy...
Pictures
Public
STAT5003_Redevelopment

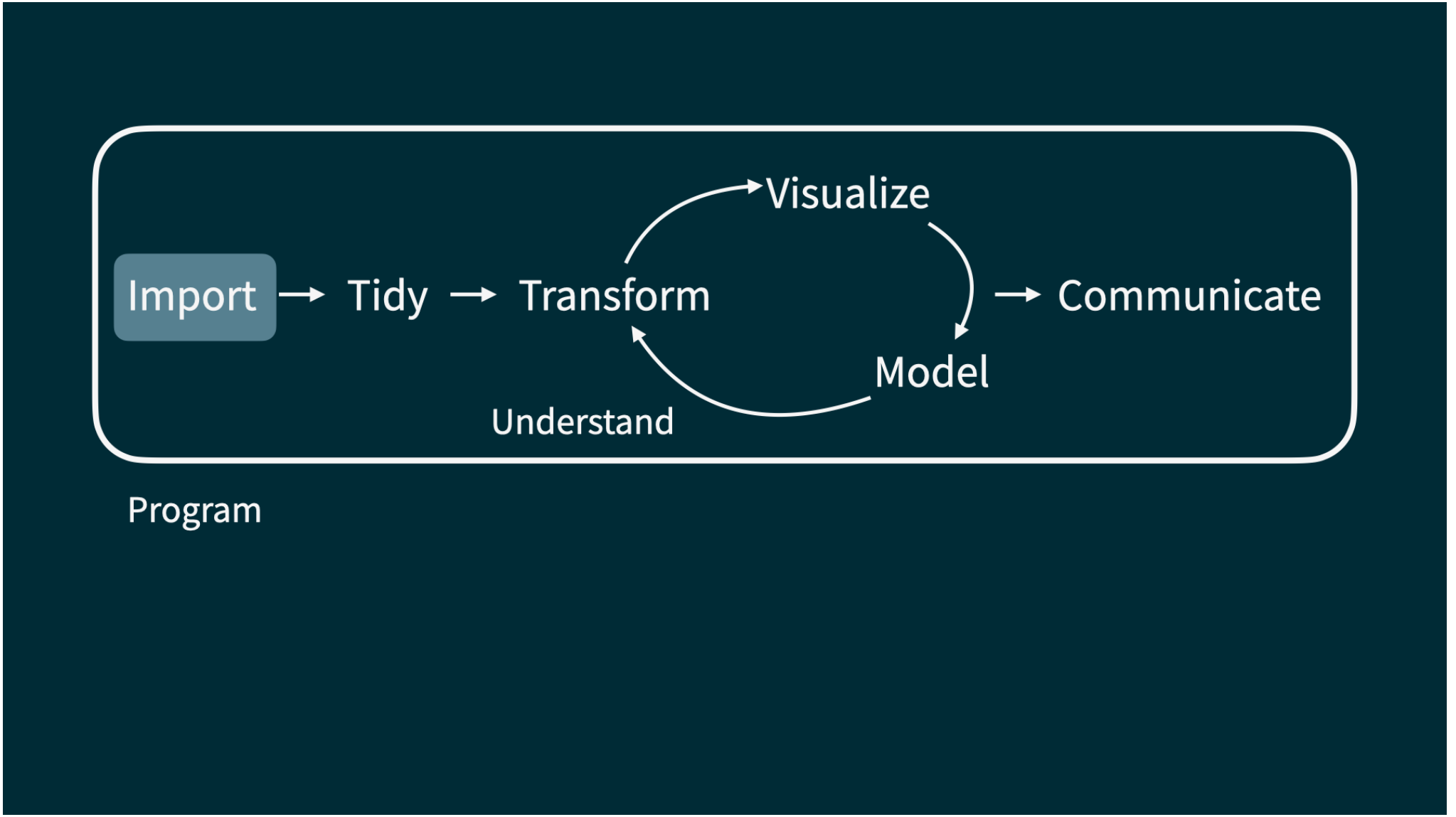
R versus RStudio

R: Engine	RStudio: Dashboard
	

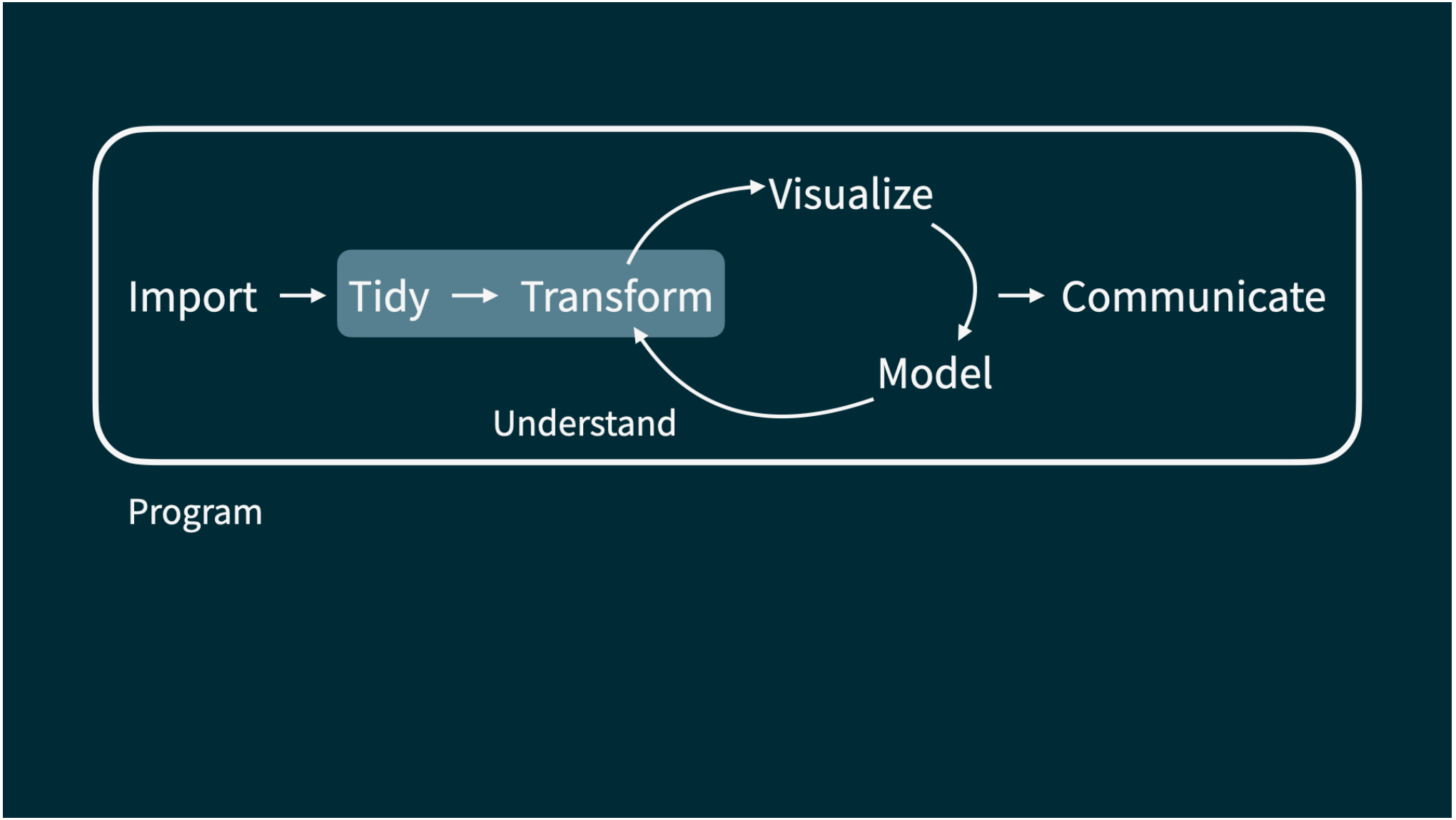
Data Science Cycle



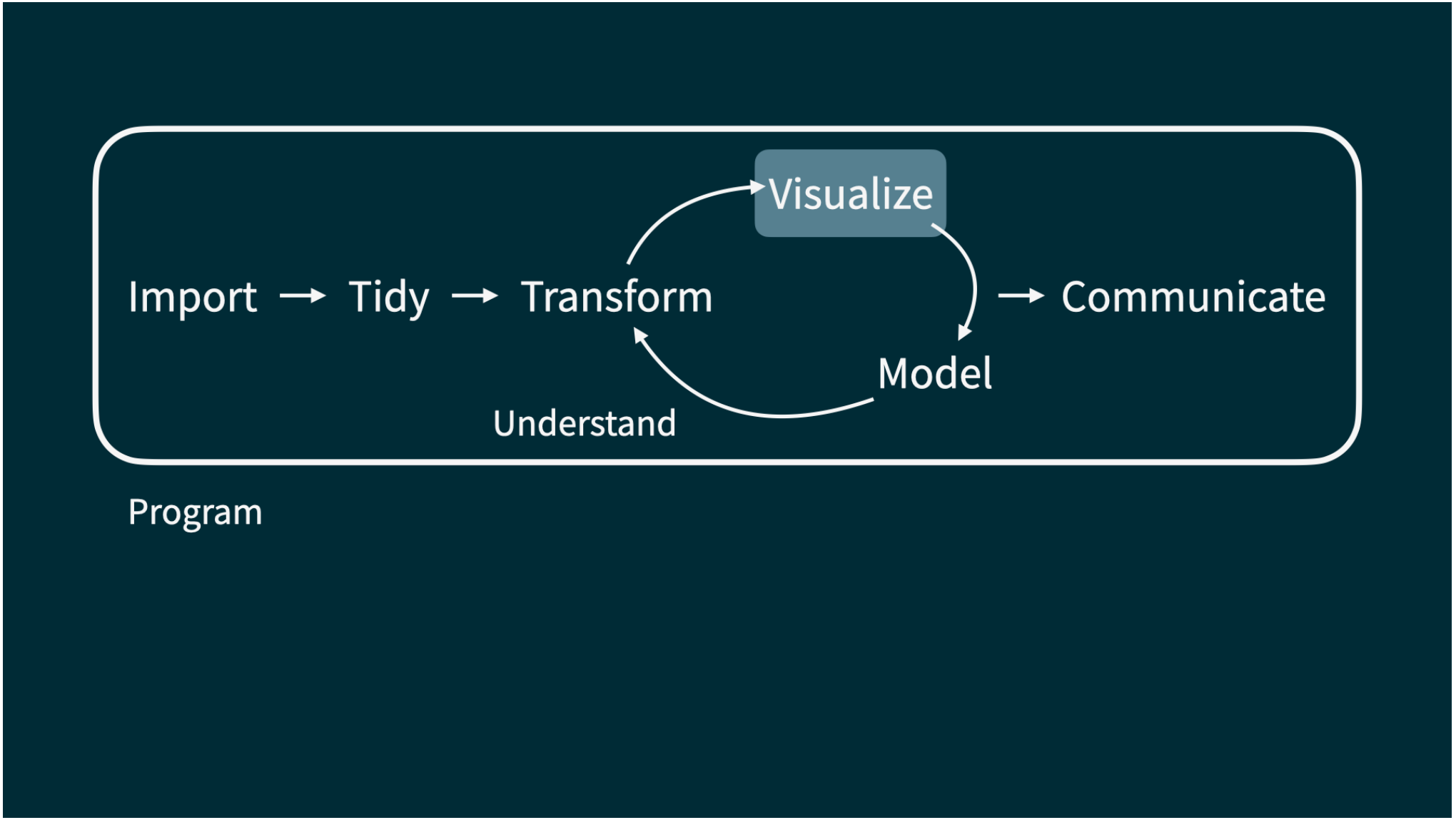
Import


Import
Tidy
Visualize
Transform
Model
Communicate
Understand
Program

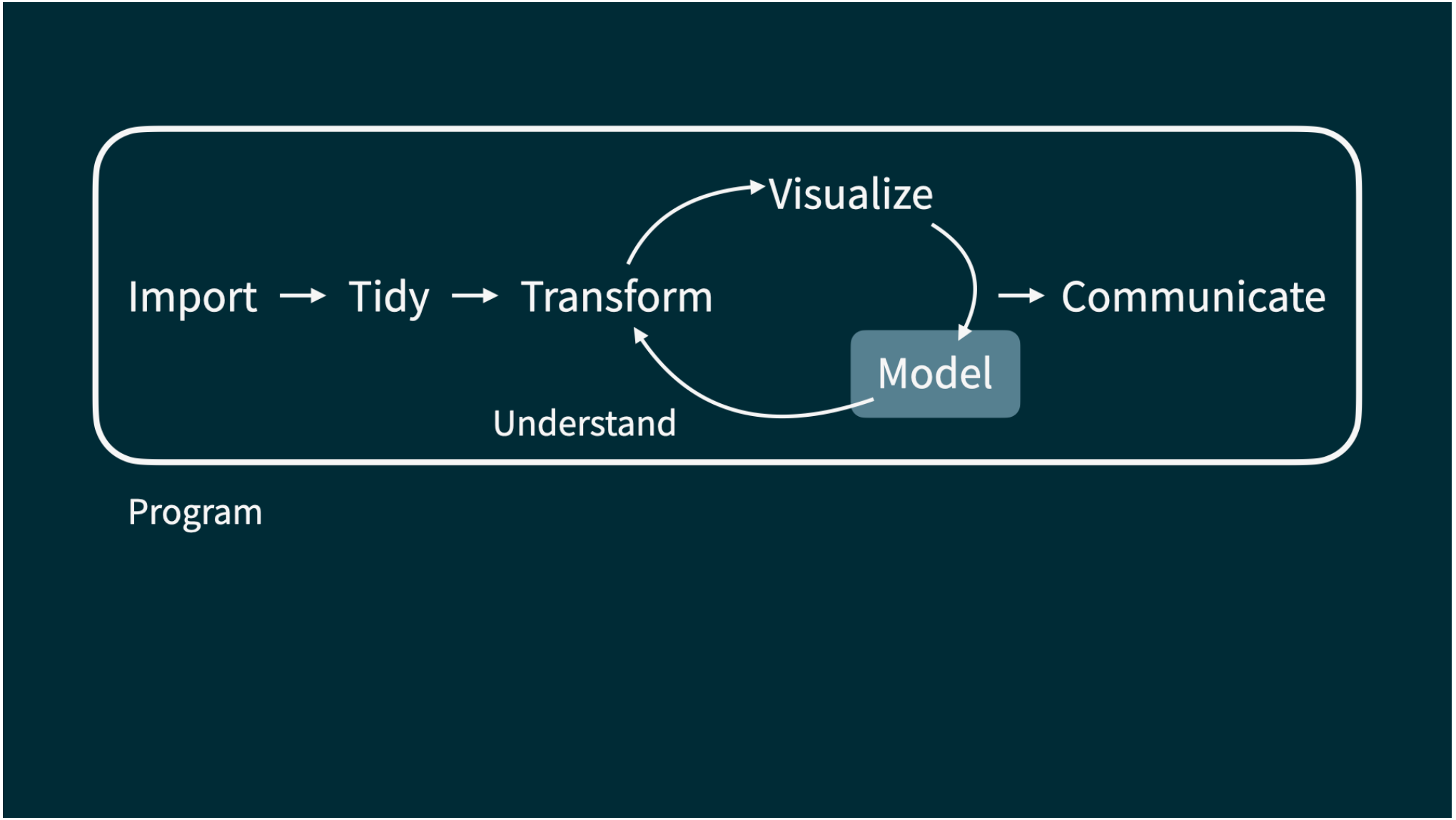
Tidy + transform



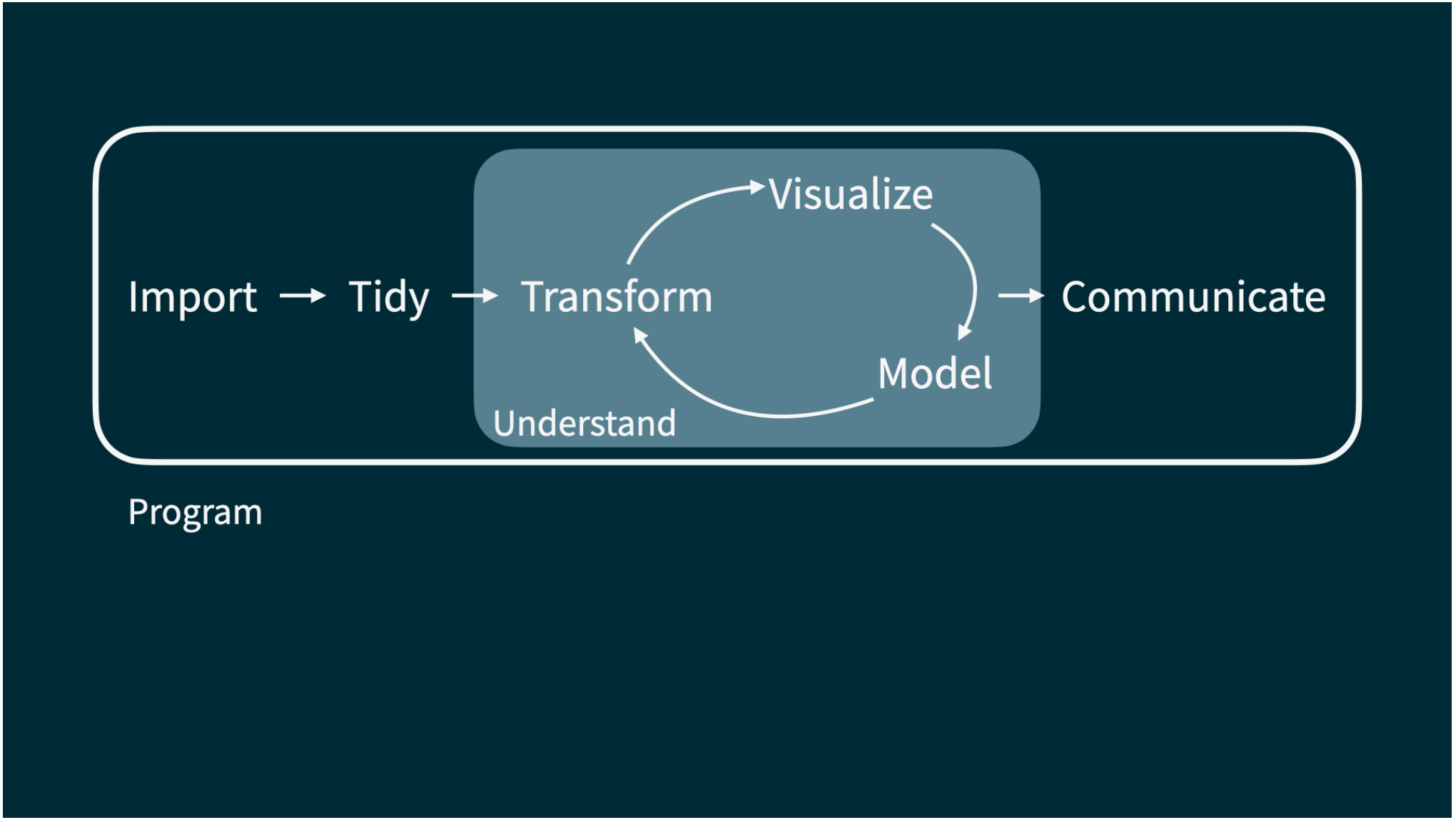
Visualize


Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program

Model


Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program

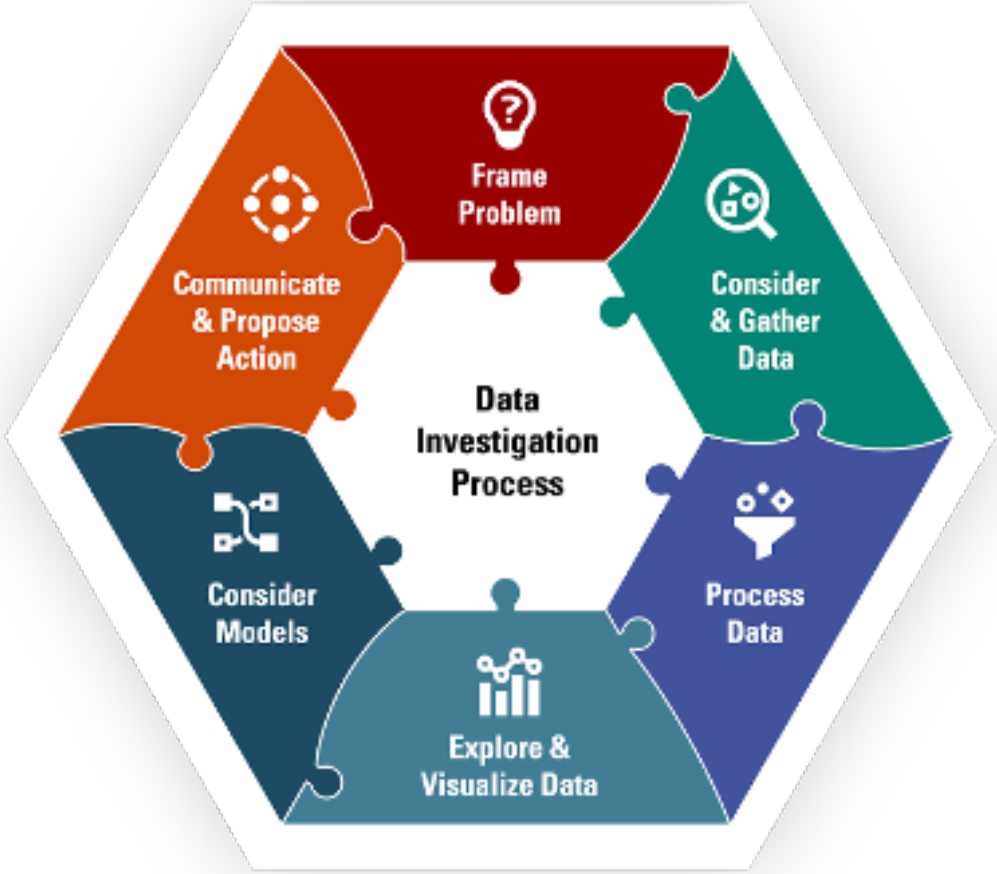
Understand


Import
Tidy
Transform
Visualize
Model
Understand
Communicate
Program

Data Story: What causes Australian road fatalities?

“Road deaths are largely predictable and preventable — a fact public health experts have sought to underscore by discouraging use of the word “accident” when it comes to road crashes.

Data Investigation Process



Data source

Frame problem

Possible research questions:

- How many road fatalities have there been so far this year, and how does it compare to last year?
- **What is the most common day for a crash?**
- Does gender affect the type of road fatality?
- What is the chance that a motorcycle rider is involved in a road fatality?

Consider and Gather Data

Data from the Australian Bureau of Statistics (ABS) (last updated Nov 2023). This dataset contains information about all road crash fatalities in Australia from 1989 to 2023. Detailed information can be found in the data dictionary.

Process Data

```r
# Read in data
data = read.csv("data/2023fatalities.csv", header=TRUE)
# Names of Variables
names(data)
```

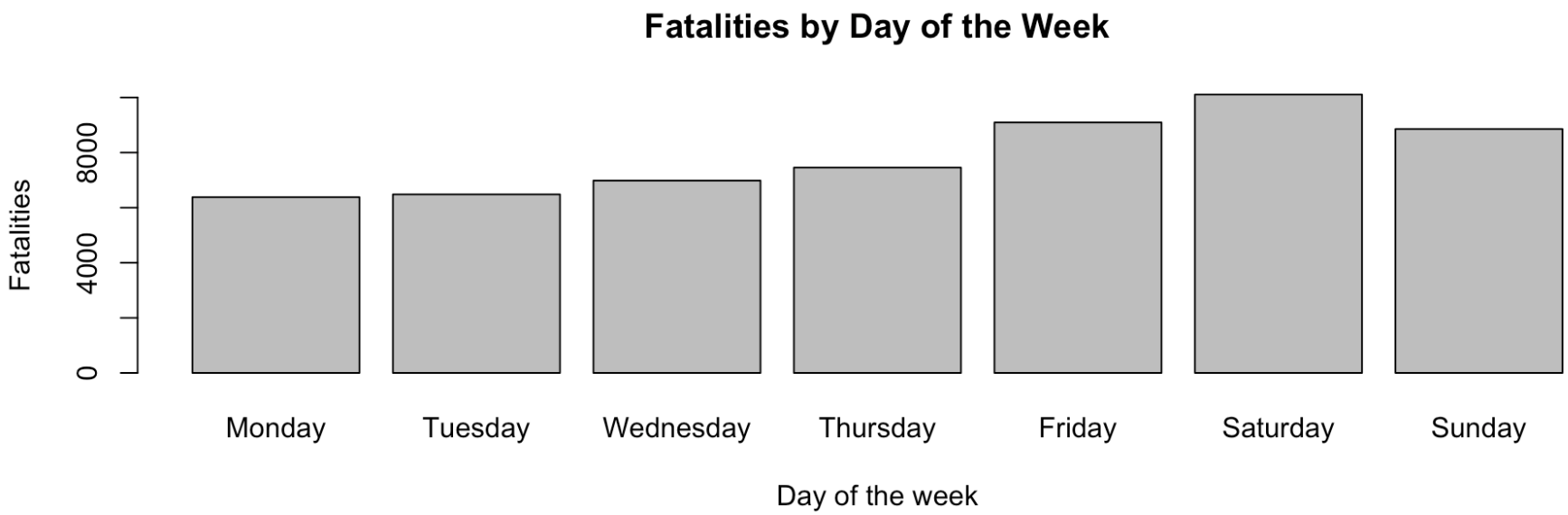
```
 [1] "Crash.ID"                      "State"
 [3] "Month"                         "Year"
 [5] "Dayweek"                       "Time"
 [7] "Crash.Type"                    "Bus.Involvement"
 [9] "Heavy.Rigid.Truck.Involvement" "Articulated.Truck.Involvement"
[11] "Speed.Limit"                   "Road.User"
[13] "Gender"                        "Age"
[15] "National.Remoteness.Areas"     "SA4.Name.2021"
[17] "National.LGA.Name.2021"        "National.Road.Type"
[19] "Christmas.Period"              "Easter.Period"
[21] "Age.Group"                     "Day.of.week"
[23] "Time.of.day"                   "X"
```

```r
# Select the DayWeek variable from the whole data frame
Dayweek = data$Dayweek

# Order days
Dayweek=factor(Dayweek,levels= c("Monday", "Tuesday","Wednesday", "Thursday",
                                 "Friday", "Saturday", "Sunday"))
table(Dayweek)
```

```
Dayweek
   Monday   Tuesday Wednesday  Thursday    Friday  Saturday    Sunday
     6382      6483      6983      7456      9094     10107      8855
```

Explore and Visualise Data

```
barplot(table(Dayweek), main = "Fatalities by Day of the Week",
        ylab = "Fatalities", xlab = "Day of the week")
```



Your turn

1. Is there an equal distribution of road fatalities across all days of the week?
2. Are weekends associated with a higher number of road fatalities compared to weekdays?

Consider Models

Goodness of fit test

Assuming road fatalities are equally likely on any give day of the week,

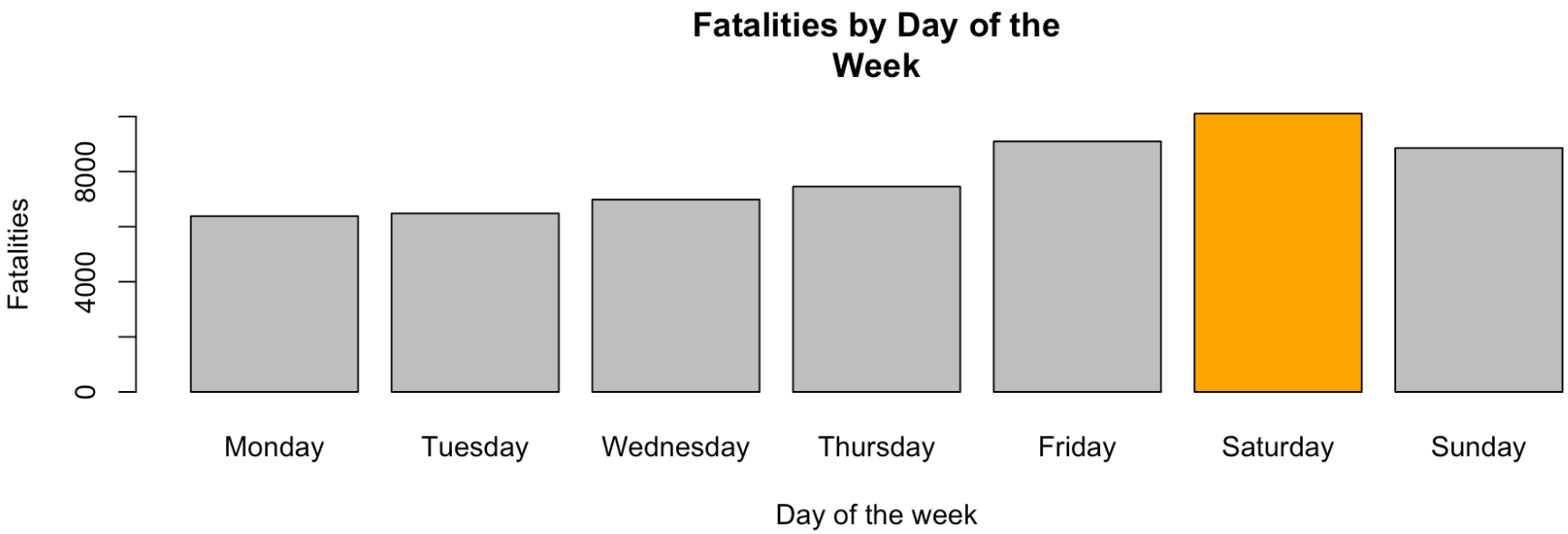
- We would like to test the hypothesis $H_0: p_1 = \cdots = p_7 = \frac{1}{7}$.
- We are interested in **any alternative that is not H_0**.
  - That is, $p_j \neq \frac{1}{7}$ for at least one $j = 1, \ldots, 7$.
  - In brief the alternative is $H_1:$ **not H_0**.
- This is an example of a **goodness of fit test** (more on this later):

```
1 chisq.test(table(Dayweek))
```

```
    Chi-squared test for given probabilities

data:  table(Dayweek)
X-squared = 1587.9, df = 6, p-value < 2.2e-16
```

Communicate and Propose Action

Given the **p-value <0.05**, we have strong evidence against the null hypothesis that (H_0) assumed that the probability of road fatalities is the same for each day of the week.



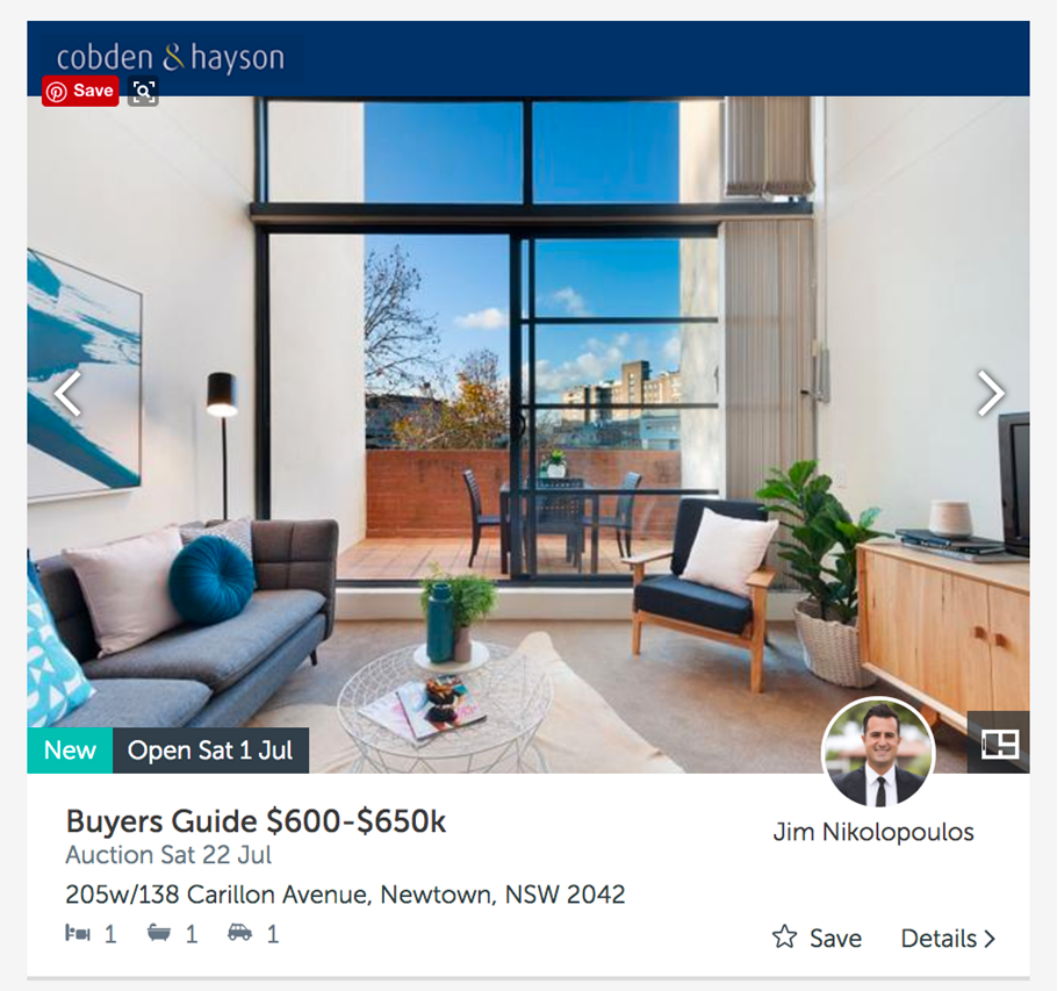
Communicate and Propose Action

> **Note**
>
> **Data insights**: it seems that Saturdays have higher road fatalities. This can help in targeting interventions more effectively.

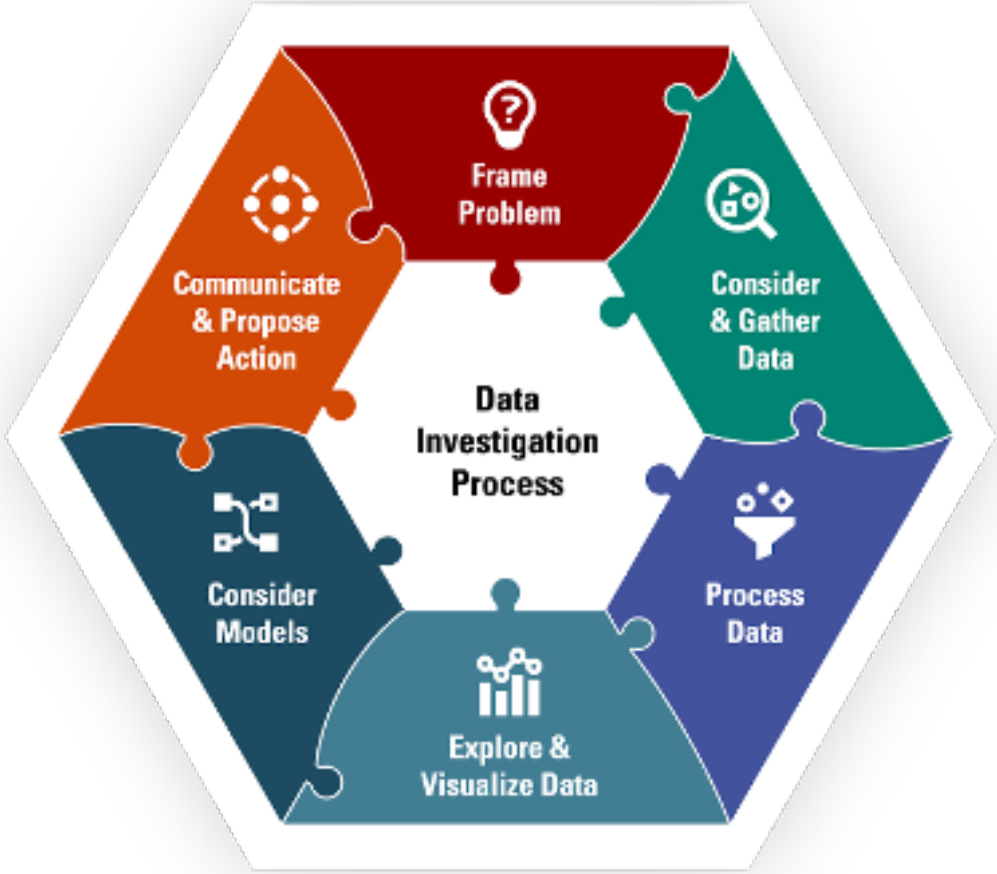
> **Note**
>
> **Future Research**: investigate the underlying causes of higher fatalities on Saturdays, which involves looking into factors such as weather conditions, traffic volume, and driver behavior.

Data story: Newtown property price

Newtown, known for its vibrant community and unique character. By analysing house prices, we can help potential buyers and investors make informed decisions.



Data Investigation Process



Data source

Frame Problem

Is the **mean house price** equal to 1.8 millions in Newtown?

Here you can find what you need to know about **Newtown, NSW**, including house prices in the area, median values, annual growth, recent sale prices, maps, a suburb profile and much more.

Suburb Insights for Houses in last 12 months

At a Glance

$1.8M	1.84%	$880pw	2.5%
Median Sale Price	Median Price Change (1 yr)	Median Rent	Median Gross Yield
(National Average $485k)	(National Average 4.30%)	(National Average $395)	(National Average 4.24%)
297th In NSW	936th In NSW	238th In NSW	867th In NSW

Consider and Collect Data 

- Data is taken from domain.com.au:
  - All properties sold in Newtown (NSW 2042) between April-June 2017.
  - The variable `Sold` has price in $1000s.

Process Data 

```r
library(tidyverse)
data <- read.csv("data/NewtownJune2017.csv",header=T)
head(data, n=2)
```

```
                 Property  Type    Agent Bedrooms Bathrooms Carspots Sold
1 19 Watkin Street Newtown House RayWhite        4         1         1 1975
2  30 Pearl Street Newtown House RayWhite        2         1         0 1250
     Date
1 23/6/17
2 23/6/17
```

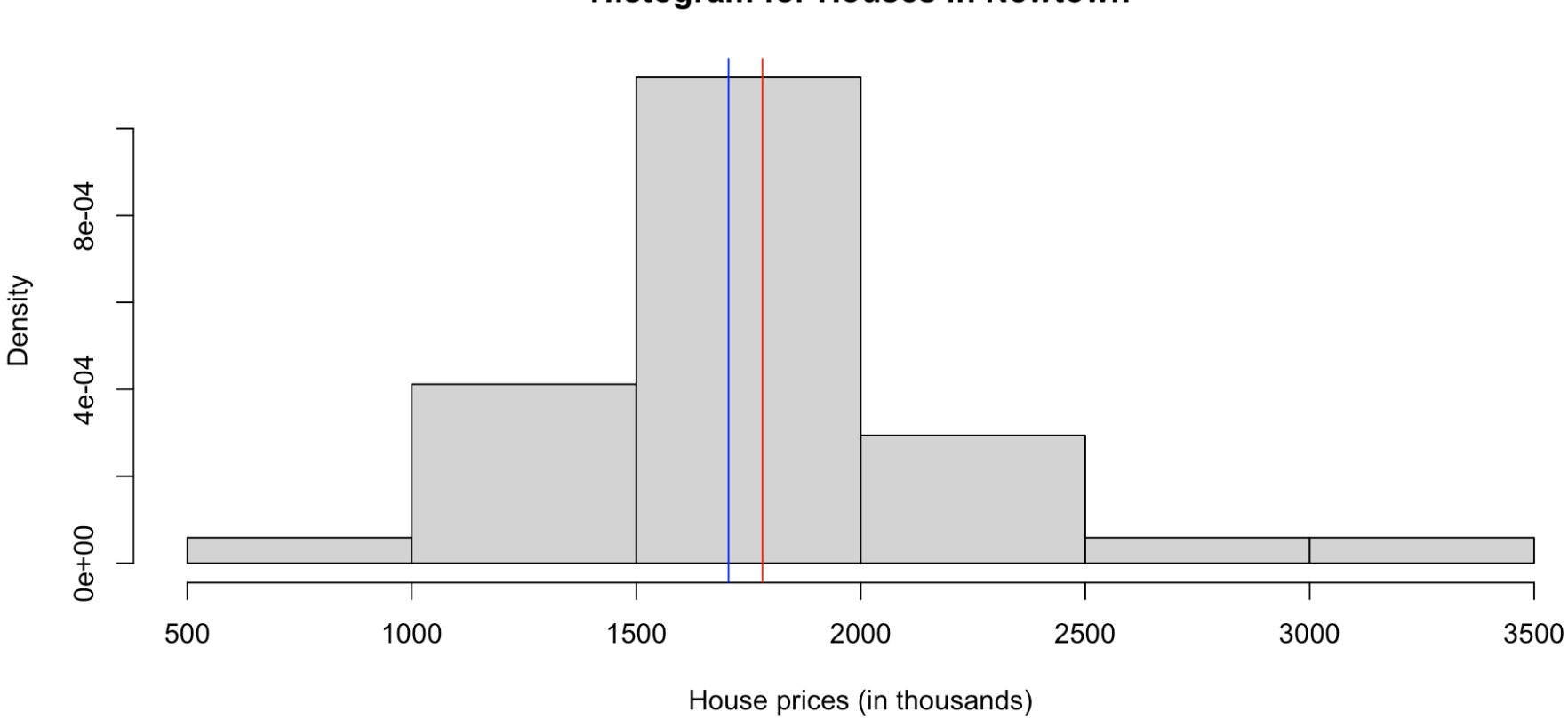
```r
## Extract observations whose Type == "House"
houses <- data |> filter(Type == "House") |> select(Sold)
```

Explore and Visualise Data

```r
houses |>  summarise(mean = mean(Sold), median = median(Sold),
                                          sd = sd(Sold))
```

```
     mean median       sd
1 1781.059 1705.5 446.6888
```

```r
hist(houses$Sold, freq = F,  xlab="House prices (in thousands)", ylab="Density",
     main="Histogram for Houses in Newtown")
abline(v = mean(houses$Sold), col = "red")
abline(v = median(houses$Sold), col = "blue")
```


Histogram for Houses in Newtown
Density
0e+00
4e-04
8e-04
500
1000
1500
2000
2500
3000
3500
House prices (in thousands)

Consider Models

One-sample t-test

Assuming that there is no signficant change in house prices in Newtown,

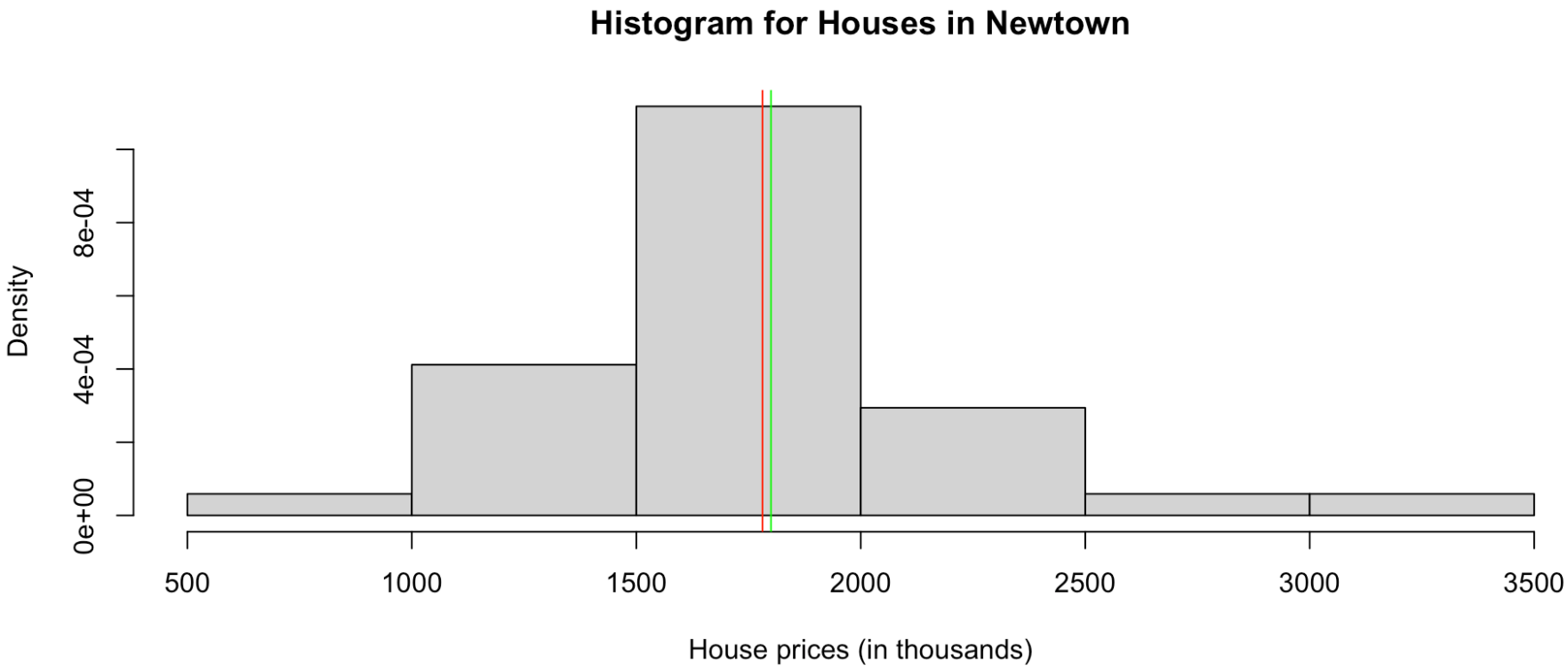
- We would like to test the hypothesis $H_0: \mu_0 = 1800$.
- We are interested in **any alternative that is not H_0**.
  - That is, $\mu_1 \neq 1800$.
  - In brief the alternative is $H_1:$ **not H_0**.
- This is an example of a **one-sample t-test** (more on this later):

```r
t.test(houses$Sold, mu = 1800, conf.level = 0.95)
```

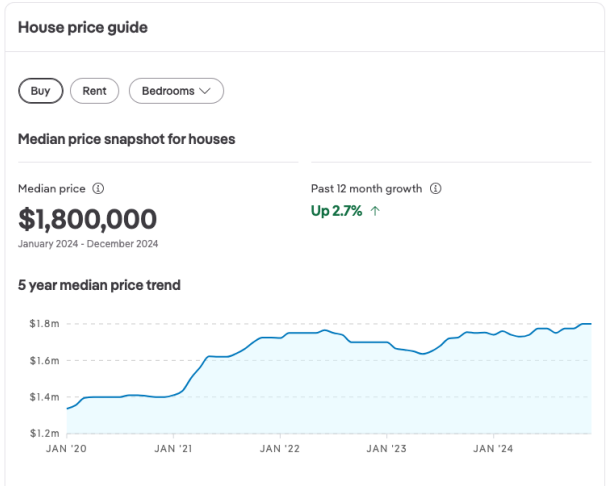
```
    One Sample t-test

data:  houses$Sold
t = -0.24725, df = 33, p-value = 0.8062
alternative hypothesis: true mean is not equal to 1800
95 percent confidence interval:
 1625.202 1936.916
sample estimates:
mean of x
 1781.059
```

Communicate and Propose Action

Given the **p-value >0.05**, we find no evidence against the null hypothesis that (H_0) suggesting that the mean house price in Newtown is $1.8 millions.



Communicate and Propose Action



Note

Data insights: It seems that house prices in Newtown remains strong and steady.

Note

Future Research: It is known that **house prices tend to rise with an increase in the number of bathrooms and the size of the land**.

Communicate and Propose Action

- Exploring the relationship between the number of bathrooms and house prices.