

# R for Data Science

<https://r4ds.had.co.nz/>

**Il-Youp Kwak, PhD**

# Objectives

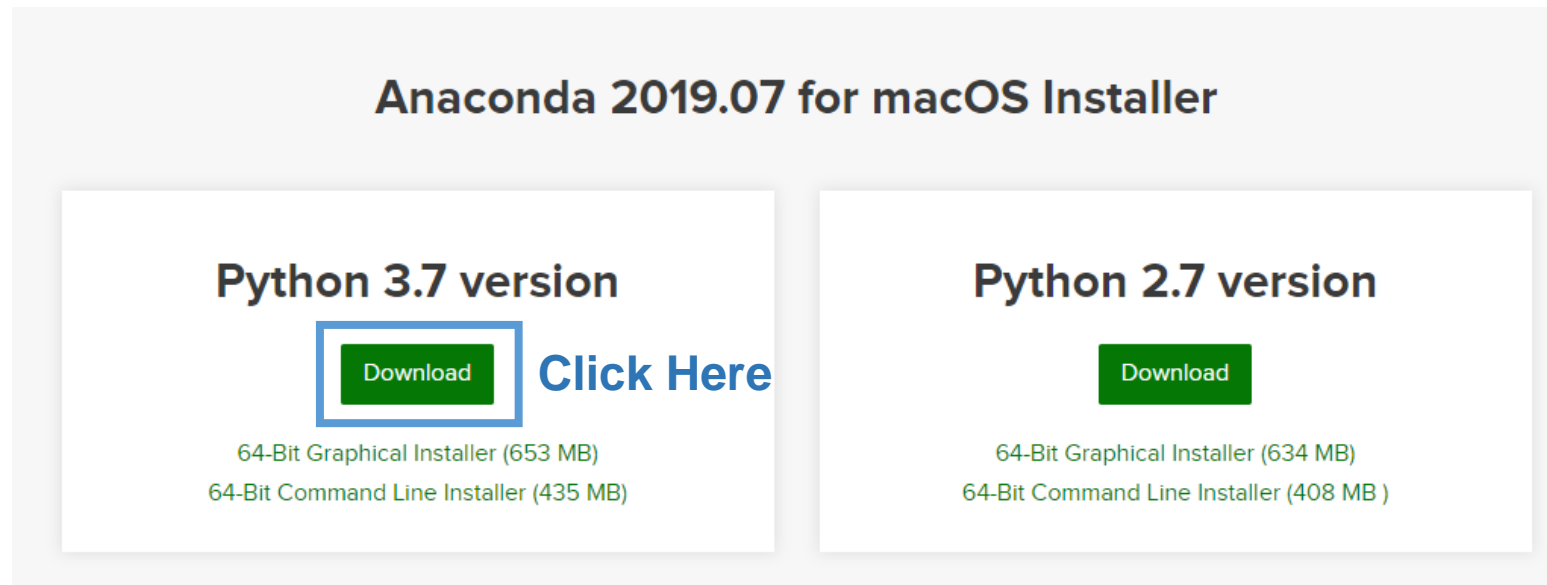
- Learn some basics in <https://r4ds.had.co.nz/>
- Especially, **dplyr** and **ggplot**

# **Try Jupyter Notebook with R**

- Easier to learn**
- Can work on both R and Python**
- Easy to read .ipynb from github unlike .Rnw**

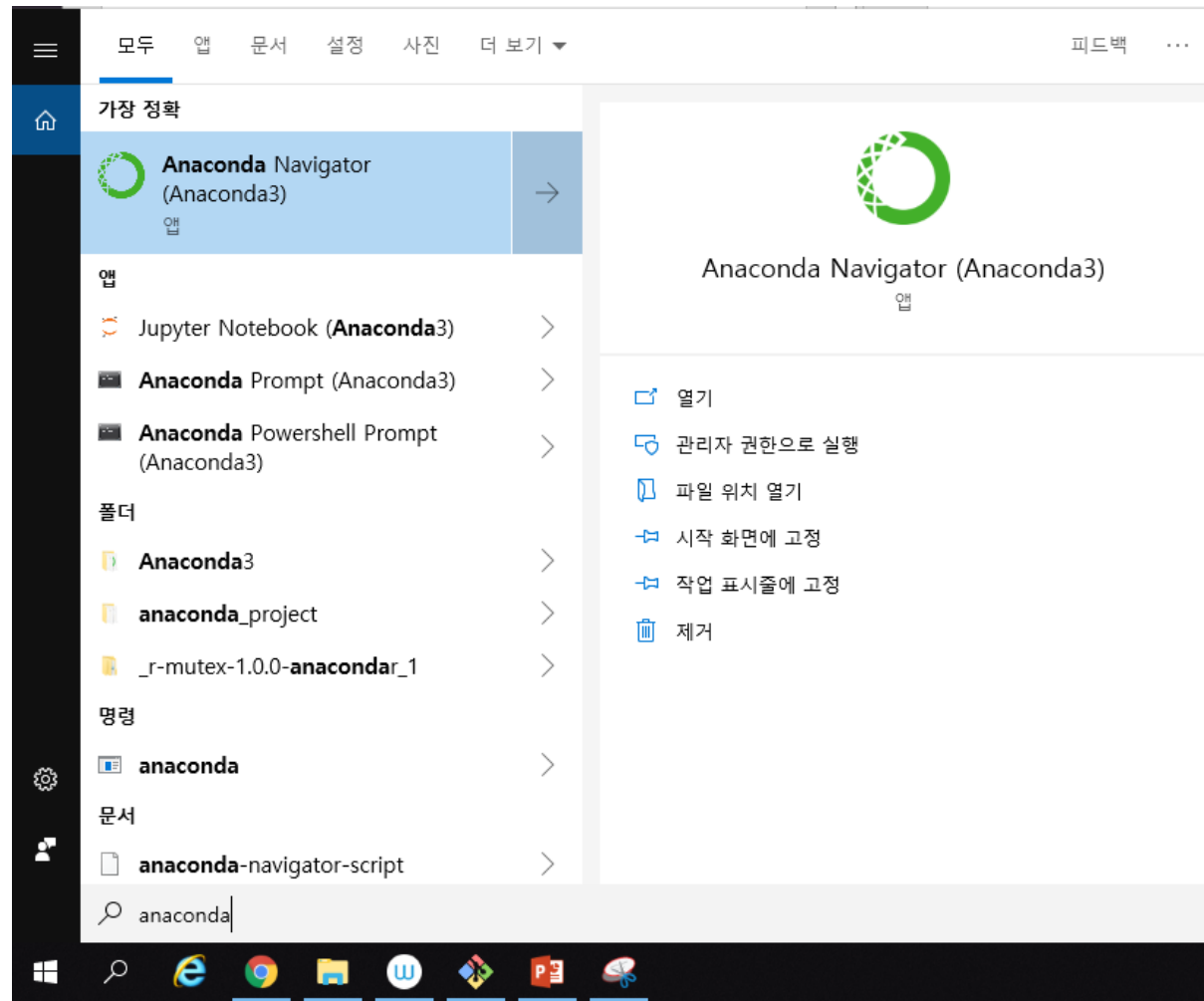
# What is Anaconda?

- **World's Most Popular Python/R Data Science Platform**
- **Downloadable:** <https://www.anaconda.com/distribution/>



# Run R from Jupyter Notebook

## - Run Anaconda Navigator



# Run R from Jupyter Notebook

Anaconda Navigator

File Help

ANACONDA NAVIGATOR

Sign in to Anaconda Cloud

Home

Environments

1. Click Environments

Community

Documentation

Developer Blog



Search Environments

base (root)

my\_env

2. Click Create



Create



Clone



Import



Remove

Installed

Channels

Update index...

Search Packages

Name Description

Version

✓	_ipyw_jlab_nb_ex...	A configuration metapackage for enabling anaconda-bundled jupyter extensions	0.1.0
✓	alabaster	Configurable, python 2+3 compatible sphinx theme.	0.7.12
✓	anaconda	Simplifies package management and deployment of anaconda	2019.07
✓	anaconda-client	Anaconda.org command line client library	1.7.2
✓	anaconda-project	Tools for managing anaconda projects	0.8.3
✓	asn1crypto	Python 2 and 3 compatible ASN.1 parser and serializer	0.24.0
✓	astroid	Static analysis tool for Python code	2.2.5
✓	astropy	Astronomy and astrophysics community Python package	3.2.1
✓	atomicwrites	A module for writing to files atomically	1.3.0
✓	attrs	Classes that automatically have attributes	19.1.0
✓	babel	Utilities to internationalize and localize python applications	2.7.0
✓	backcall	Specifications for callback functions passed in to an api	0.1.0
✓	backports	Backports of python 3.x features to python 2.x	1.0
✓	backports.functoo...	Backport of functools.lru_cache from python 3.3 as published at activestate.	1.5
✓	backports.os	Backport of new features in python's os module	0.1.1
✓	backports.shutil_g...	A backport of the get_terminal_size function from python 3.3's shutil.	1.0.0
✓	backports.tempfile	Backport of python 3.x's tempfile module	1.0
✓	backports.weakref	Backport of new features in python's weakref module	1.0.post1
✓	beautifulsoup4	Python library designed for screen-scraping	4.7.1
✓	bitarray	Efficient arrays of booleans -- c extension	0.9.3

273 packages available

3. Setup R environment

Create new environment

Name: my\_env

Location: C:\Users\CAU\Anaconda3\envs\my\_env

Packages: ☒ Python 3.6

☒ R r

Cancel

Create

you from the drudgery of implementing object protocols (aka dunder methods).

# Run R from Jupyter Notebook

Anaconda Navigator

File Help

ANACONDA NAVIGATOR

Sign in to Anaconda Cloud

Home

Environments

Learning

Community

Documentation

Developer Blog



Search Environments

Installed

Channels

Update index...

Search Packages

base (root)

my\_env

Open Terminal

Open with Python

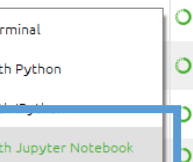
Open with Jupyter Notebook

Name

Description

Version

Open with Jupyter Notebook



	Attrs is the python package that will bring back the joy of writing classes by relieving you from the drudgery of implementing object protocols (aka dunder methods).		1.0.0
	Specifications for callback functions passed in to an api		19.1.0
	Easy, python-based html sanitizing tool		0.1.0
			3.1.0
	Python package for providing mozilla's ca bundle.		2019.6.16
	Cross-platform colored terminal text.		0.4.1
	Better living through python with decorators.		4.4.0
	Xml bomb protection for python stdlib modules		0.6.0
	Discover and load entry points from installed packages.		0.3
	Ipython kernel for jupyter		5.1.2
	Ipython: productive interactive computing		7.8.0
	Vestigial utilities from ipython		0.2.0
	An autocompletion tool for python that can be used for text editors.		0.15.1
	An easy to use stand-alone template engine written in pure python.		2.10.1
	An implementation of json schema validation for python		3.0.2
	Jupyter protocol implementation and client libraries		5.3.1
	Core common functionality of jupyter projects.		4.5.0
	A modern and easy-to-use crypto library.		1.0.16
			1.9.10
			1.0.6

221 packages available

Create Clone Import Remove

# Run R from Jupyter Notebook

 jupyter

QuitLogout

FilesRunningClusters

Select items to perform actions on them.

☐ 0 /

UploadNew↺

Name

Notebook:  
Python 3

R

Other:  
Text File  
Folder  
Terminal

<input type="checkbox"/> 3D Objects	
<input type="checkbox"/> Anaconda3	
<input type="checkbox"/> Contacts	
<input type="checkbox"/> Desktop	
<input type="checkbox"/> Documents	
<input type="checkbox"/> Downloads	16분 전
<input type="checkbox"/> Favorites	7일 전
<input type="checkbox"/> Links	7일 전
<input type="checkbox"/> Music	7일 전
<input type="checkbox"/> OneDrive	7일 전
<input type="checkbox"/> Pictures	3일 전
<input type="checkbox"/> R	20일 전
<input type="checkbox"/> Saved Games	7일 전
<input type="checkbox"/> Searches	7일 전
<input type="checkbox"/> Videos	7일 전
<input type="checkbox"/> miktex-console.lock	20일 전

Open with R



# What is ggplot ?

- is a package for data visualization
- Use grammar of graphics

Alternatives: R base functions

# Aesthetic mapping

- **ggplot( aes( x = , y= , color= , size= , alpha= , shape=) )**

**Just try it**

<https://r4ds.had.co.nz/data-visualisation.html#aesthetic-mappings>

# Facets

- **Split your plot into facets (facet\_wrap, facet\_grid)**

**Just try it**

<https://r4ds.had.co.nz/data-visualisation.html#facets>

# Geometric objects

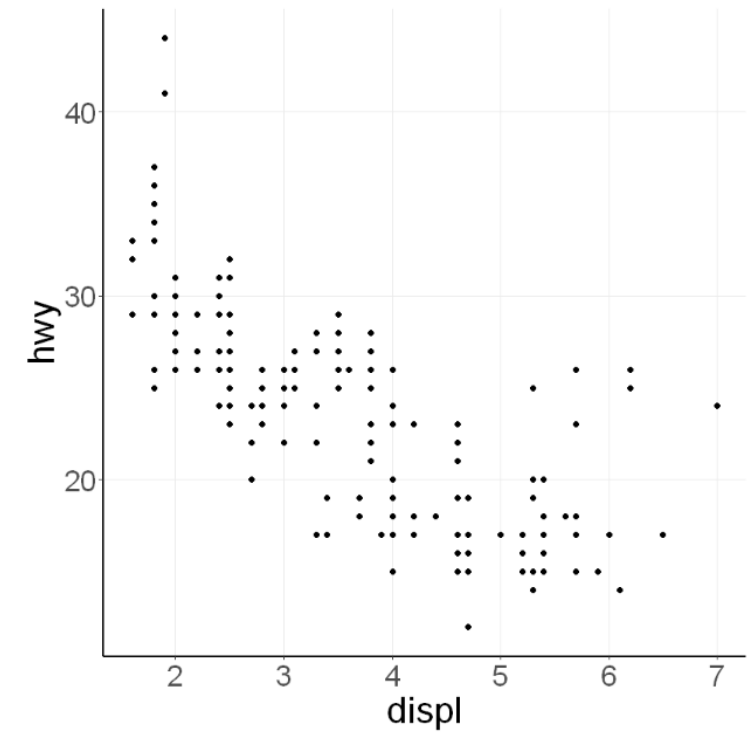
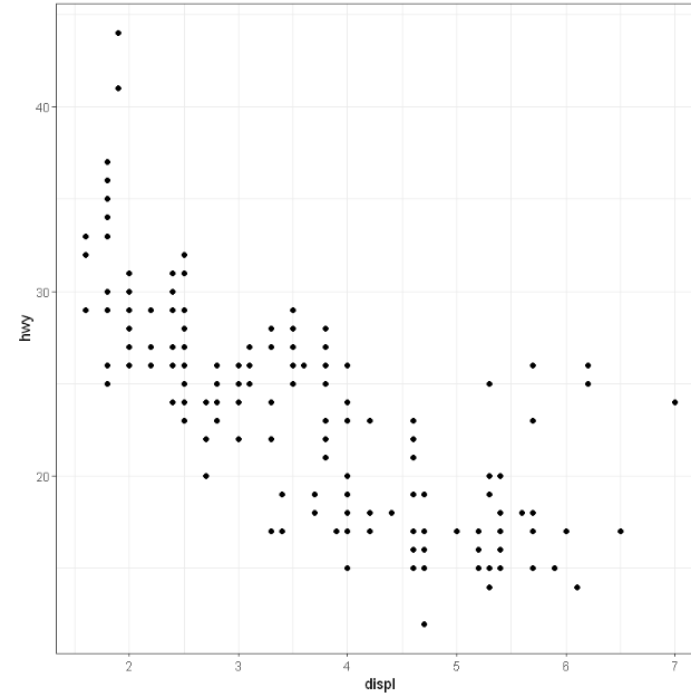
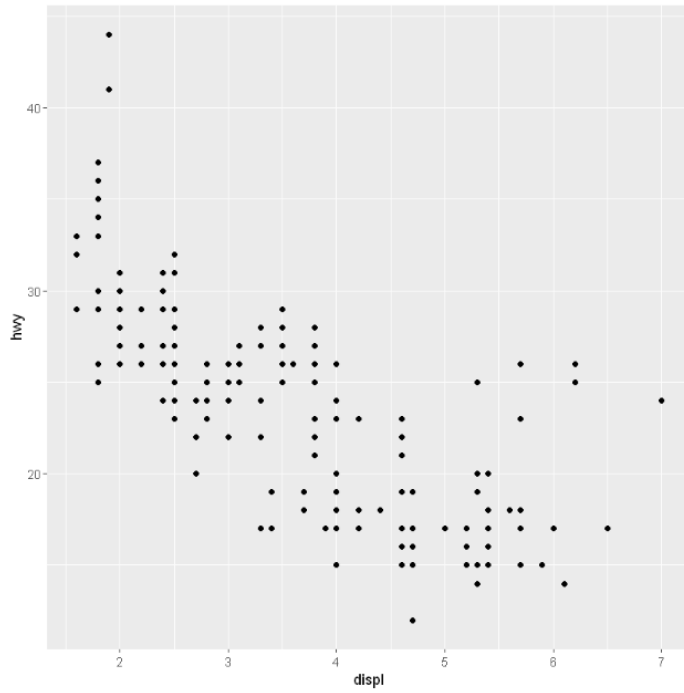
- **geom** stand for geometrical objects
- **geom\_point**, **geom\_smooth**, **geom\_bar**, **geom\_violin**, **geom\_abline**, etc

**Just try it**

<https://r4ds.had.co.nz/data-visualisation.html#geometric-objects>

# Theme

- What will you choose?

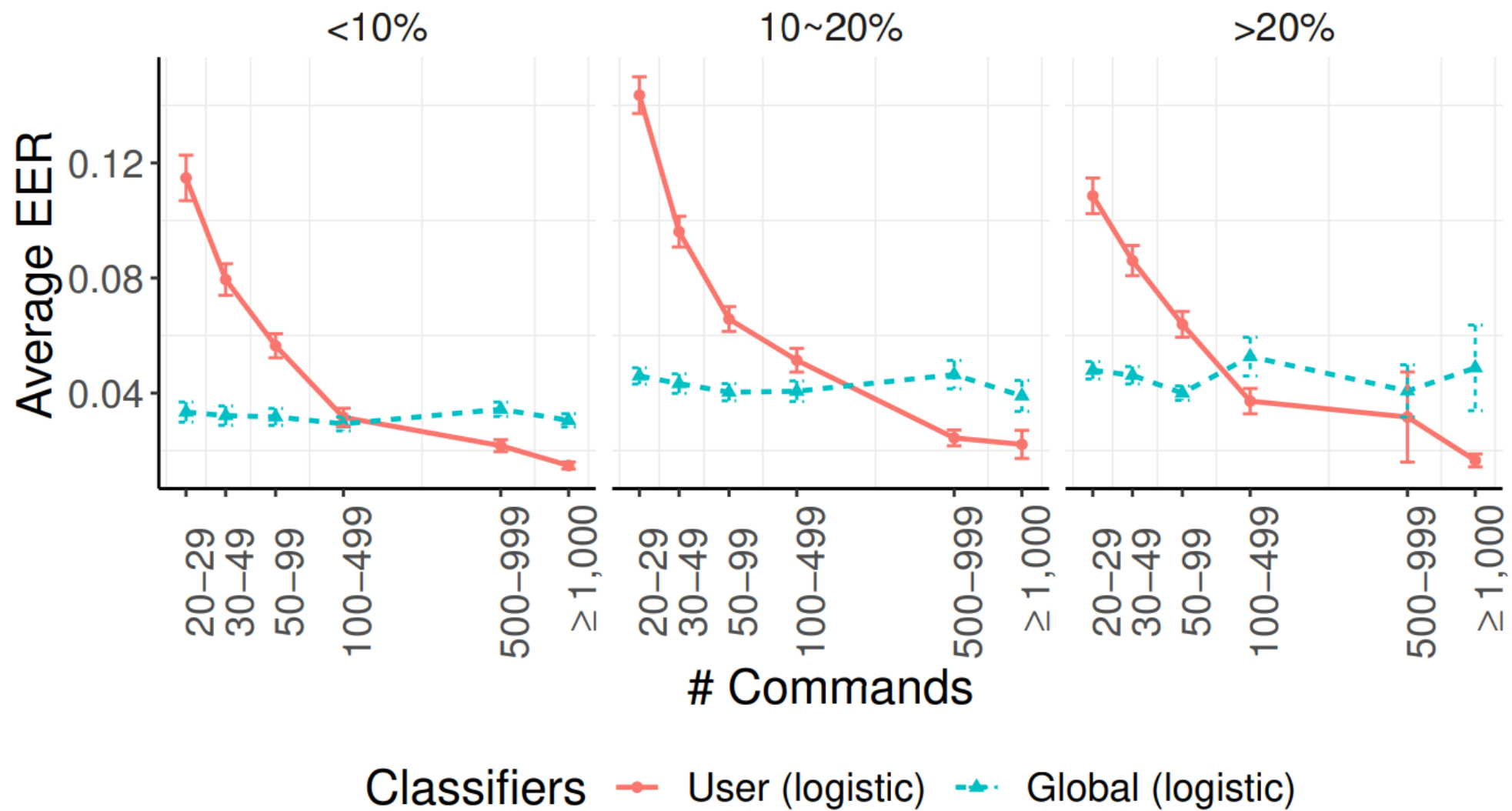


# Theme

- Theme that I like

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy)) + theme_bw() +  
  theme(axis.line = element_line(size = .8, colour = "black"),  
        panel.grid.minor = element_blank(),  
        panel.border = element_blank(),  
        text = element_text(size = 25)  
  )
```

# Example



# Use `postscript()` or `tiff()` for quality

```
postscript(file = "test.eps", width=7, height=7)
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) + theme_bw() +
  theme(axis.line = element_line(size = .8, colour = "black"),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        text = element_text(size = 25)
  )
dev.off()
```



# **ggplot Practice**

# What is dplyr ?

- **is a package for data manipulation**
- **Functions are coded in C++**
- **Fast and efficient**

Alternatives: data.table package, R base functions

# **filter()**

- **Return a subset of the rows**

**Just try it**

<https://r4ds.had.co.nz/transform.html#filter-rows-with-filter>

# arrange()

- **Reorders the rows according to single or multiple variables**

**Just try it**

<https://r4ds.had.co.nz/transform.html#arrange-rows-with-arrange>

# **select()**

- **Return a subset of the columns**

**Just try it**

<https://r4ds.had.co.nz/transform.html#select>

# mutate()

- **Add columns from existing data**

**Just try it**

<https://r4ds.had.co.nz/transform.html#add-new-variables-with-mutate>

# summarize()

- Produce summary statistic for each group (using `group_by()` )

**Just try it**

<https://r4ds.had.co.nz/transform.html#grouped-summaries-with-summarise>

# Improve readability using pipe operators

`x %>% f(y)` turns into `f(x, y)`, and `x %>% f(y) %>% g(z)` turns into `g(f(x, y))`.

```
by_dest <- group_by(flights, dest)
delay <- summarise(by_dest,
  count = n(),
  dist = mean(distance, na.rm = TRUE),
  delay = mean(arr_delay, na.rm = TRUE)
)
delay <- filter(delay, count > 20, dest != "HNL")
```

```
delays <- flights %>%
  group_by(dest) %>%
  summarise(
    count = n(),
    dist = mean(distance, na.rm = TRUE),
    delay = mean(arr_delay, na.rm = TRUE)
  ) %>%
  filter(count > 20, dest != "HNL")
```



# Improve readability using pipe operators

```
bop(  
  scoop(  
    hop(foo_foo, through = forest),  
    up = field_mice  
  ),  
  on = head  
)
```

```
foo_foo %>%  
  hop(through = forest) %>%  
  scoop(up = field_mice) %>%  
  bop(on = head)
```

# dplyr Practice

<https://r4ds.had.co.nz/transform.html>

# What is tidyr ?

- Represent the same data multiple ways

```
table1
#> # A tibble: 6 x 4
#>   country      year cases population
#>   <chr>      <int> <int>      <int>
#> 1 Afghanistan 1999     745  19987071
#> 2 Afghanistan 2000    2666  20595360
#> 3 Brazil       1999   37737  172006362
#> 4 Brazil       2000   80488  174504898
#> 5 China        1999  212258 1272915272
#> 6 China        2000  213766 1280428583
```

```
table2
#> # A tibble: 12 x 4
#>   country      year type      count
#>   <chr>      <int> <chr>      <int>
#> 1 Afghanistan 1999 cases         745
#> 2 Afghanistan 1999 population 19987071
#> 3 Afghanistan 2000 cases         2666
#> 4 Afghanistan 2000 population 20595360
#> 5 Brazil       1999 cases         37737
#> 6 Brazil       1999 population 172006362
#> # ... with 6 more rows
```

```
table3
#> # A tibble: 6 x 3
#>   country      year rate
#>   * <chr>      <int> <chr>
#> 1 Afghanistan 1999 745/19987071
#> 2 Afghanistan 2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

```
table4a # cases
#> # A tibble: 3 x 3
#>   country `1999` `2000`
#>   * <chr>      <int> <int>
#> 1 Afghanistan     745     2666
#> 2 Brazil          37737    80488
#> 3 China           212258   213766
```

```
table4b # population
#> # A tibble: 3 x 3
#>   country `1999` `2000`
#>   * <chr>      <int> <int>
#> 1 Afghanistan 19987071 20595360
#> 2 Brazil      172006362 174504898
#> 3 China       1272915272 1280428583
```

# gather()

- **gather those columns into a new pair of variables**

**Just try it**

<https://r4ds.had.co.nz/tidy-data.html#gathering>

# spread()

- opposite of gathering

**Just try it**

<https://r4ds.had.co.nz/tidy-data.html#spreading>

**Thank you!**  
**Q & A**