

# WELCOME!

- Before we get started:
  - Login to [r.umhb.edu](https://r.umhb.edu)
    - Use your UMHB email address and password
  - *File > New File > R Script*
  - Make yours look like this



The screenshot shows the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar contains icons for saving, running, and other functions. The main editor window shows a blank R script file named 'Untitled1'. The console window on the right displays the R version 3.5.0 (2018-04-23) -- "Joy in Playing" and the R Foundation for Statistical Computing copyright notice. The environment window at the bottom shows the Global Environment, which is empty. The file explorer on the right shows the Home directory with files .Rhistory, R, and R.backup.

```
R version 3.5.0 (2018-04-23) -- "Joy in Playing"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (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.

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.

> |
```

Name	Size	Modified
.Rhistory	13.6 KB	Aug 30, 2018, 1:46 PM
R		
R.backup		

# Introduction to R

Basics of Data Manipulation, Visualization, and Analysis

---

**Aaron R. Baggett, Ph.D.**

University of Mary Hardin-Baylor

PSYC 2316: Statistics for the Social Sciences

September 12, 2018



# A Gentle Introduction to R

Basics of Data Manipulation, Visualization, and Analysis

---

**Aaron R. Baggett, Ph.D.**

University of Mary Hardin-Baylor

PSYC 2316: Statistics for the Social Sciences

September 12, 2018



**WHAT IS R?**

# WHAT IS R?

- R is:
  - A powerful, flexible statistics and data software program
  - Free and open source
  - Surging in adoption worldwide
    - 1 of 3 statistics and data programming languages in the top 20<sup>[1]</sup>
  - Able to read *any* data file

[1] TIOBE Programming Index, <http://bit.ly/2x4TQGh>

# WHO USES R?



# WHO ELSE USES R?

You do!



**R YOU READY?**



# LET'S GET STARTED

- R is like a big calculator with a huge memory
- One advantage:
  - We can store input inside **names** or **objects**
- Let's try it

# LET'S GET STARTED

- In your R Script type the following:

```
a <- 2 * 5  
b <- 0:10  
c <- a + b
```

- Send all three lines to the R Console to run a, b, and c
  - Hint: Highlight and click  Run

# NYC FLIGHT DATA, 2013

# NYC FLIGHTS

- Most of the time, we want to work with real data sets using powerful tools
- In your script, add this code and run it to the console:

```
# Load tidyverse package  
library(tidyverse)  
  
# Read in nycflights data  
nycflights <- url("http://bit.ly/nyc_flights")  
load(nycflights)  
  
# View nycflights data in viewer  
View(nycflights)
```

# nycflights Glimpse

	month	dep_time	dep_delay	arr_time	arr_delay	carrier	dest
1	6	940	15	1216	-4	VX	LAX
2	5	1657	-3	2104	10	DL	SJU
3	12	859	-1	1238	11	DL	LAX
4	5	1841	-4	2122	-34	DL	TPA
5	7	1102	-3	1230	-8	9E	ORF
6	1	1817	-3	2008	3	AA	ORD
---	---	---	---	---	---	---	---
32730	1	706	36	909	22	EV	IND
32731	10	752	-8	921	-28	9E	PIT
32732	7	812	-3	1043	8	DL	LAS
32733	9	1057	-1	1319	-19	UA	IAH
32734	10	844	56	1045	60	B6	CHS
32735	3	1813	-3	1942	-23	UA	CLE

# NYC FLIGHTS

- Flight delays are always a hassle
- Let's examine all departure delays (dep\_delay)
- First, a little tutorial:

```
# General framework for summarizing  
data %>%  
  summarize(  
    object_name = mean(outcome),  
    object_name = sd(outcome))
```

# NYC FLIGHTS

- Flight delays are always a hassle
- Let's examine all departure delays (dep\_delay)
- Ready?

```
# Calculate mean and SD departure delays  
nycflights %>%  
  summarize(  
    mean_dd = mean(dep_delay),  
    sd_dd = sd(dep_delay))
```

# NYC FLIGHTS

- On average, flights departing NYC airports in 2013 were delayed by about 13 minutes ( $SD = 40$  minutes).

```
# Calculate mean and SD departure delays
```

```
nycflights %>%  
  summarize(  
    mean_dd = mean(dep_delay),  
    sd_dd = sd(dep_delay))
```

```
##      mean_dd      sd_dd  
## 1 12.70515 40.40743
```



# **EXPLAINING DEPARTURE DELAYS**

# VARIABILITY IN DEPARTURES

- What factors might explain variability in departure delays?

# VARIABILITY IN DEPARTURES

- What factors might explain variability in departure delays?
  - Weather
  - Time of year
  - Etc.

# VARIABILITY IN DEPARTURES

- Let's reexamine departure delays
- This time, we will account for the time of year (month)
- What do you think we will find?

# VARIABILITY IN DEPARTURES

- Let's add two more lines to our previous section

```
# General framework for summarizing with grouping  
data %>%  
  group_by(grouping_var) %>%  
  summarize(  
    object_name = mean(outcome),  
    object_name = sd(outcome)) %>%  
  arrange(desc(sorted_var))
```

# VARIABILITY IN DEPARTURES

- Let's add two more lines to our previous section

```
# General framework for summarizing with grouping
data %>%
  group_by(grouping_var) %>%
  summarize(
    object_name = mean(outcome),
    object_name = sd(outcome)) %>%
  arrange(desc(sorted_var))
```

# VARIABILITY IN DEPARTURES

- To what extent to departure delays vary by month of travel?
- Ready?

```
# Departure delays by month
month_dd <- nycflights %>%
  group_by(month) %>%
  summarize(
    mean_dd = mean(dep_delay),
    sd_dd = sd(dep_delay)) %>%
  arrange(desc(mean_dd))
```

# VARIABILITY IN DEPARTURES

- Longer travel delays appear to occur during the mid-summer and Christmas season.

```
## # A tibble: 12 x 3
##   month mean_dd sd_dd
##   <int>   <dbl> <dbl>
## 1     7    20.8   47.8
## 2     6    20.4   53.5
## 3    12    17.4   43.0
## 4     4    14.6   43.4
## 5     3    13.5   40.3
## 6     5    13.3   38.3
## 7     8    12.6   39.2
## 8     2    10.7   33.1
## 9     1    10.2   42.4
## 10     9     6.87  35.3
## 11    11     6.10  27.6
## 12    10     5.88  29.4
```



# **VISUALIZING DEPARTURE DELAYS**

# VISUALIZING DELAY PATTERNS

- Let's visualize these the pattern of the departure delays
- First, a little tutorial:

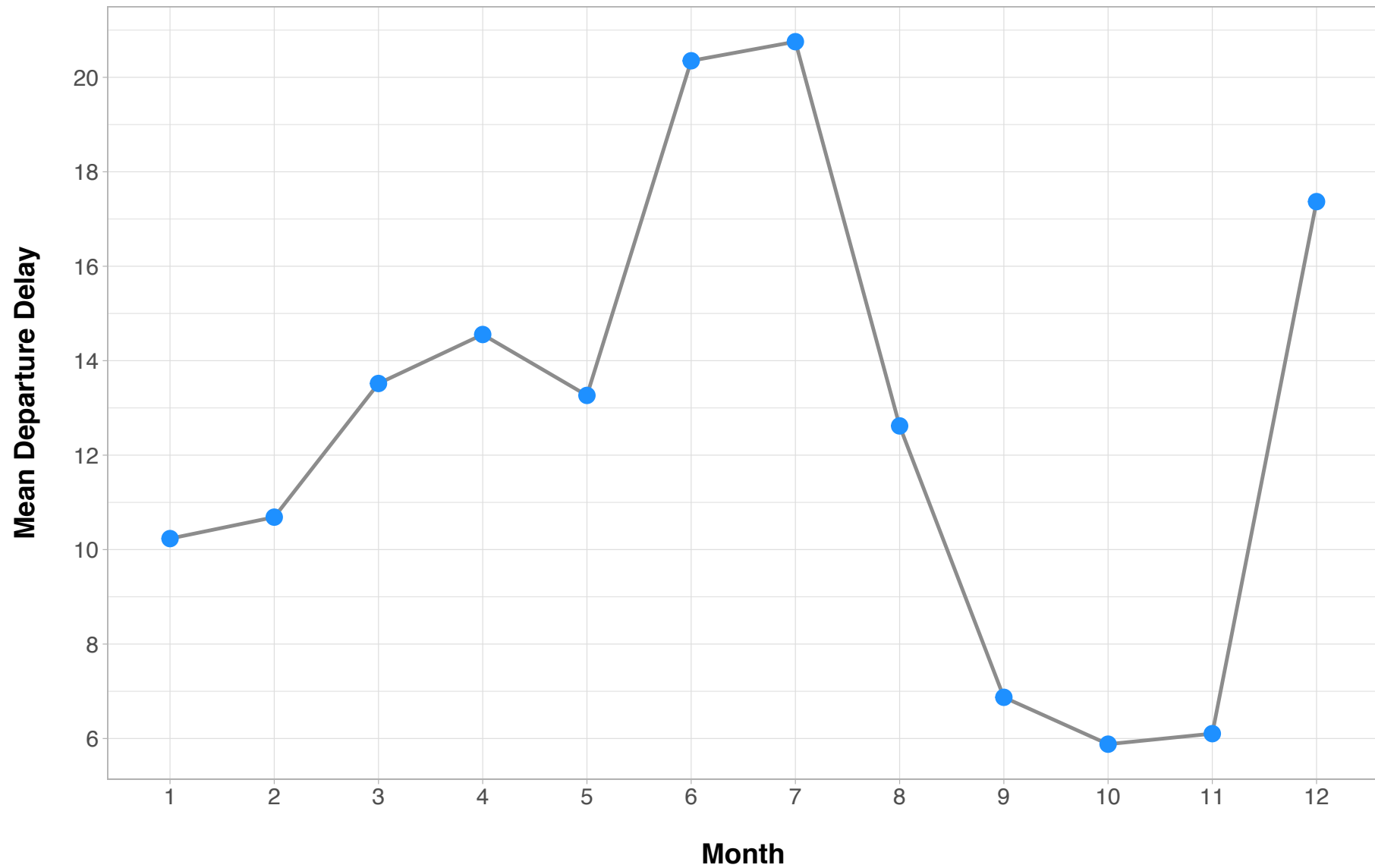
```
ggplot(data = data,  
  aes(x = factor, y = outcome, group = 1)) +  
  geom_point() +  
  geom_line()
```

# VISUALIZING DELAY PATTERNS

- Let's visualize these the pattern of the departure delays
- Ready?

```
ggplot(data = month_dd,  
  aes(x = month, y = mean_dd, group = 1)) +  
  geom_point() +  
  geom_line()
```

# VISUALIZING DELAY PATTERNS



**RECAP**

# RECAP

- Steep learning curve at first
- Flexibility in and power in variety of tools
- Makes analysis reproducible
- Adoption surging

**QUESTIONS?**

# GET IN TOUCH

**Aaron R. Baggett, Ph.D.**

Department of Psychology

[abaggett@umhb.edu](mailto:abaggett@umhb.edu)

Ext. 4553