STAT 100: Week 3

Ricky's Section

Introductions and Attendance

Introduction: Name

<u>**Question of the Week</u>**: What is one word to describe how you're feeling? Try not to repeat words!</u>

Important Reminders



https://docs.google.com/forms/d/e/1FAIpQLSfKv FGvsooqm-IvtxKx3Vf6bBzSJE2jamK1gklAzL6Nk XE8w/viewform

Join our Slack: #section-d003

https://join.slack.com/share/enQtNzY3NDUxNT QxMzkwOS1lNzVhZWQxN2ZmYWMyYThlMzIyY mM4ZmNiMTJmOWViZGE2MzQ5NWVkMTM2Y WY4MWRhMWZiZDAxOWZlZDYxYmYx

Content Review: Week 2

Grammar of Graphics

https://drive.google.com/file/d/1nAVN8_qi1vWC wllXjgat7MukSv5Wgpqm/view?usp=sharing

Choosing the Right Graph

https://drive.google.com/file/d/1T1rovHS3l1rapz RH6fGNclS2fn6lu7gs/view?usp=sharing



- Make sure your code isn't running off the screen in your PDF
- Hit "Return" to start a new line
 - Best to do this after commas (,) and plus signs (+)





🗯 Firefox File Edit Vie	w History Bookmarks Tools Window Help	n 🗧 🚺 🔿
🗕 🕘 💼 📥 STAT 10	10 Teachii 🗙 🚦 STAT 100 Sectio: X 📃 STAT 100 Sectio: X 🔺 STAT 100 - Rem: X 🌼 Course Roster: S X	💥 All Content - STA 🗙 🔀 Pro
$\leftrightarrow \rightarrow \mathbf{G}$	C A 🕫 https://posit.cloud/spaces/534915/content/8756398	ជ
🕅 Notion 🔲 Calendar 附 Main Ma	il 附 School Mail 🚍 Main Docs 🚍 School Docs 🔺 Main Drive 🔺 School Drive 👿 Harvard 🍀 Canvas 🛐 S	mbolab 👖 Gradescope 💓 Posit C
≥ posit Cloud	STAT 100 Fall 2024 / Problem Set 2	
Spaces	File Edit Code View Plots Session Build Debug Profile Tools Help	
(Vour Workspace	🔹 🔹 🚽 🚰 👘 🚔 👘 Go to file/function	
Tour workspace	🕏 pset02_stat100_fall2024.Rmd × 📃 colleges × 📃 colleges_moderate_removal × 📃 colleges_>>> 🗇	Environment History Conne
STAT 100 Fall 2024 Harvard University	🕼 🖓 💭 🔄 🖓 🔍 🖋 Knit 🔹 💮 🔹 🦉 🖓 🖓 🖓 🖓	😅 🔒 📑 Import Dataset 👻 🤇
The ford Only Clarky	Source Visual a Outline	R 👻 💼 Global Environment 👻
New Space	337 a) Let's examine one measure of mobility rate: the percentage of students with	Data
	parents in the bottom income quintile who ended up in the top income quintile	O colleges 1285
	average annual cost of attendance: customize the color and transparency of the	Colleges_aggressi 279 c
Learn	`geom`. Describe what you see.	Ocolleges_light_re 1285
@ Cuida	338	O colleges_moderate 1284
Guide	340	
What's New	<pre>341 ggplot(data = colleges, mapping = aes(y = sticker_price_2013, x = sat_avg_2013,</pre>	Files Plats Performent Hall
	<pre>color = tier_name)) +</pre>	Files Plots Packages Help
Recipes	$342 geom_point(alpha = 0.5) + \\ 343 geom_point(alpha = 0.5) + \\ 343 $	New Folder V New Blank Fil
	144 labs(x = "Average SAT in 2013", y = "Sticker Price in 2013", color = "Tier	Cloud > project
Cheatsheets	Name", title = "Colleges in America")	A Name
	345	
	346	. Rhistory
Help	5 m	data
dh e se e se	Lings Arms	project.Rproj
'III' Current System Status		pset02_stat100_fall202
Resit Community		pset02_stat100_fall202
	" viet first	
	(38:5:232m`geom smooth()` using formula = 'v ~ x'[39m]	
Info	A Warning: [38;5;232mRemoved 341 rows containing non-finite outside the scale	
	range (`stat_smooth()`).[39m	
\$ Plans & Pricing	Warning: [38;5;232mRemoved 341 rows containing missing values or values	
_	outside the scale range (`geom_point()`).[39m	
Terms and Conditions	Collogos in America	
	337:349 EProblem 4 C R Markdown C	
	Console 🗗 🗖	

<mark>Clean Code</mark>

🗯 Firefox File Edit Vie	w History Bookmarks Tools Window Help	🕞 🍝 💴 🚳 (
🔍 🔍 💼 🖾 STAT 10	0 Teachii X 🚦 STAT 100 Sectioi X 📃 STAT 100 Sectioi X 💧 STAT 100 - Remi: X 🎲 Course Roster: S X	💥 All Content - STA 🗙 🛛 🗮 Prot
$\leftrightarrow \rightarrow \mathbf{C}$	♡ 읍 ☞ https://posit.cloud/spaces/534915/content/8756398	☆
🕅 Notion 🧾 Calendar M Main Mai	ii 附 School Maii 📃 Main Docs 📃 School Docs 🔺 Main Drive 💧 School Drive 👿 Harvard 🌼 Canvas 🛐 Syr	nbolab 📊 Gradescope 💓 Posit C
posit Cloud 💿	STAT 100 Fall 2024 / Problem Set 2	
Spaces	File Edit Code View Plots Session Build Debug Profile Tools Help	
(Your Workspace	🔍 • 🚰 • 🔒 🚔 🍐 Go to file/function 🔛 • Addins •	
	🗢 pset02_stat100_fall2024.Rmd × 🔹 colleges × 🔹 colleges_moderate_removal × 🔹 colleges_>>> 👝 🗇	Environment History Conne
Harvard University	(iii) I I I I I A Knit + ⊙ + 10 N I I I I I I I I I I I I I I I I I I	🚰 🔒 📑 Import Dataset 👻 🛛
	Source Visual @Outline	R 👻 📑 Global Environment 👻
New Space	337 a) Let's examine one measure of mobility rate: the percentage of students with parents in the bottom income quintile who ended up in the top income quintile	Data
	(`mr_kq5_pq1`). Create a plot that shows the association between `mr_kq5_pq1` and	Colleges 1285
	average annual cost of attendance; customize the color and transparency of the	Colleges_aggressi 279 d
Learn	'geom'. Describe what you see.	Colleges_light_re 1285
Ø Guide	339 x ```{n} ◎ ≍ ▶	Colleges_moderate 1284
	340	
What's New	<pre>341 ggplot(data = colleges, mapping = aes(y = sticker_price_2013,</pre>	Files Plots Packages Helr
	$x = sat_avg_2013,$	New Folder New Blank Eil
Recipes	343 acom_point(alpha = 0.5) +	
	345 geom_smooth(method = "lm", se = FALSE) +	A Name
Cheatsheets	<pre>346 labs(x = "Average SAT in 2013",</pre>	1
	347 y = "Sticker Price in 2013", 348 color = "Tier Nome"	Phistony
	349 title = "Colleges in America")	
Help	350	
dly commencement of the	351	project.kproj
-in- current system status	332 *	Diseto2_stat100_fail202
<section-header> Posit Community</section-header>		pset02_stat100_tall202
Info		
\$ Plans & Pricing	 [38;5;232m`geom_smooth()`using formula = 'y ~ x'[39m Warning: [28:5:32mBernaued 241 roug containing non finite outside the scale 	
_	warming. [56;5;2:22mkemoved 541 rows containing non-inite outside the scale range (`stat_smooth()`) [39m	
Terms and Conditions	Warning: (20:5-222)mDamound 241 rows containing missing values or values	
	348:11 S Chunk 12 C R Markdown C	
	Console	

Reading Code in English

<u>In R</u>: ggplot(data = colleges, mapping = aes(y = sticker_price_2013, x = sat_avg_2013, color = tier_name)) + geom_point(alpha = 0.5)

<u>In English</u>: Create a **plot** using the "colleges" **dataset**. Map to the **aesthetic** of **y-direction** the **variable** of sticker price, map to the **aesthetic** of **x-direction** the **variable** of average SAT, and map to the **aesthetic** of **color** the **variable** tier name. Use **points** as the **geom**, and make **alpha/transparency** equal to 0.5.



Coding: In Words

<u>In R</u>: ggplot(data = colleges, mapping = aes(y = sticker_price_2013, x = sat_avg_2013, color = tier_name)) + geom_point(alpha = 0.5) + geom_smooth(method = "lm", se = FALSE) + labs(x = "Average SAT in 2013", y = "Sticker Price in 2013", color = "Tier Name", title = "Colleges in America")



<u>In English</u>: Create a **plot** using the "colleges" **dataset**. Map to the **aesthetic** of **y-direction** the **variable** of sticker price, map to the **aesthetic** of **x-direction** the **variable** of average SAT, and map to the **aesthetic** of **color** the **variable** tier name. Use **points** as the **geom**, and make **alpha/transparency** equal to 0.5.

Additionally, use the geom of smooth/line to create lines of best fit. Additionally, insert labels to the x-axis, y-axis, legends, and title.

Content Review: Week 3

<mark>Data Joins</mark>

- Use to join datasets
 via a key (variable
 to link the 2
 datasets)
- Left, Right, Inner, and Full



<mark>Left Join</mark>

- left_join(houses, students, join_by("name" == "house"))
- Combine 2 datasets via key, keeping all original observations from LEFT-HAND dataset while adding matching observations from RIGHT-HAND dataset

<pre>students ## id conc house sleep ## 1 001 CPB Winthrop 7 ## 2 002 HDRB Currier 8 ## 4 004 Econ Mather 9 ## 5 005 Psych Pfoho 6 ## 6 006 Stat Winthrop 7 ## 7 007 IB Pfoho 8 houses ## 1 Dunster 1930 River East 0 merges 7 1000 ## 3 Currier 1970 Quad 1 merges 7 1000 ## 4 Mather 1970 River East 1 merges 7 1000 for metch but kept left_join(houses, students, join_by("name" == "house")) ## name built area id conc sleep ## 1 Dunster 1930 River East </pre> NA NA T ## 2 Winthrop 1931 River West 001 CPB 7 s ## 3 Winthrop 1931 River West 003 Stat 8 merges ## 4 Winthrop 1931 River West 003 Stat 8 merges ## 4 Winthrop 1931 River West 006 Stat 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9 ##													
<pre>## id conc house sleep ## 1 001 CPB Winthrop 7 ## 2 002 HDRB Currier 8 ## 3 003 Stat Winthrop 8 ## 4 004 Econ Mather 9 ## 5 005 Psych Pfoho 6 ## 6 006 Stat Winthrop 7 ## 7 007 IB Pfoho 8 houses ## 1 Dunster 1930 River East 0 merces 7 1000 ## 3 Currier 1970 Quad 1 merces 7 1000 ## 4 Mather 1970 River East 1 merces 7 1000 cetor ictt_gon()</pre>		S	tu	dents									
<pre>## 1 0001 CPB Winthrop 7 ## 2 002 HDRB Currier 8 ## 3 003 Stat Winthrop 8 ## 4 004 Econ Mather 9 ## 5 005 Psych Pfoho 6 ## 6 006 Stat Winthrop 7 ## 6 006 Stat Winthrop 7 ## 1 Dunster 1930 River East 0 methods 7 lows ## 2 Winthrop 1931 River West 3 method 7 lows ## 3 Currier 1970 Quad 1 method 7 lows ## 4 Mather 1970 River East 1 method 7 lows Cetor /cftorn() ## name built area id conc sleep ## 1 Dunster 1930 River East <na> NA T ## 2 Winthrop 1931 River West 001 CPB 7 \$## 3 Winthrop 1931 River West 003 Stat 8 method \$## 4 Winthrop 1931 River West 003 Stat 8 ## 4 Winthrop 1931 River West 006 Stat 7 ## 6 Mather 1970 River East 004 Econ 9 </na></pre>		- #	#	id	cond	: 1	house	sleep					
<pre>## 2 Vol2 Infinite State Winthrop 8 ## 4 004 Econ Mather 9 ## 4 004 Econ Mather 9 ## 5 005 Psych Pfoho 6 ## 6 006 Stat Winthrop 7 ## 7 007 IB Pfoho 8 houses ## 1 Dunster 1930 River East 0 merches 7 1 res ## 2 Winthrop 1931 River West 5 merches 7 1 res ## 4 Mather 1970 River East 1 merches 7 1 res ## 4 Mather 1970 River East 1 merches 7 1 res for /// ## 4 Mather 1970 River East 1 merches 7 1 res for /// ## 4 Mather 1970 River East 1 merches 7 1 res for /// ## 4 Mather 1970 River East 1 merches 7 1 res for /// ## 4 Mather 1970 River East 1 merches 7 1 res for /// ## 4 Mather 1970 River East 1 merches 7 1 res for /// ## 4 Mather 1970 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 1 res for /// ## 1 Dunster 1930 River East 1 merches 7 res for /// ## 1 Dunster 1930 River East 1 merches 7 res for // ## 2 Winthrop 1931 River West 001 CPB 7 for ## 3 Winthrop 1931 River West 003 Stat 8 merches ## 4 Winthrop 1931 River West 006 Stat 7 for ## 5 Currier 1970 Quad 002 HDRB 8 for ## 6 Mather 1970 River East 004 Econ 9 for ## 6 Mather 1970 River East 004 Econ 9 for ## 6 Mather 1970 River East 004 Econ 9 for ## 6 Mather 1970 River East 004 Econ 9 for ## 6 Mather 1970 River East 004 Econ 9 for ## 6 Mather 1970 River East 004 Econ 9 for ## 6 Mather 1970 River East 004 Econ 9 for ## 6 Mather 1970 River East 004 Econ 9 for ## 6 Mather 1970 River East 004 Econ 9 for # # 6 Mather 1970 River East 004 Econ 9 for # # 6 Mather 1970 River East 004 Econ 9 for # #</pre>		#	#	1 001	UDBE	S Win	rrior						
<pre>## 4 004 Econ Mather 9 ## 5 005 Psych Pfoho 6 ## 6 006 Stat Winthrop 7 ## 7 007 IB Pfoho 8 houses ## 1 Dunster 1930 River East 0 merdes 7 1 merdes ## 2 Winthrop 1931 River West 3 merdes 7 3 merdes ## 3 Currier 1970 Quad 1 merdes 7 1 merdes ## 4 Mather 1970 River East 1 merdes 7 1 merdes ## 4 Mather 1970 River East 1 merdes for any join_by("name" == "house")) ## name built area id conc sleep ## 1 Dunster 1930 River East <na> NA T ## 2 Winthrop 1931 River West 001 CPB 7 S ## 3 Winthrop 1931 River West 003 Stat 8 meres ## 4 Winthrop 1931 River West 006 Stat 7 ## 6 Mather 1970 River East 004 Econ 9</na></pre>	1	#	#	3 003	Stat	Wint	throp	8					
<pre>## 5 005 Psych Pfoho 6 ## 6 006 Stat Winthrop 7 ## 7 007 IB Pfoho 8 houses ## 1 Dunster 1930 River East 0 merces - 1 row ## 2 Winthrop 1931 River West 3 merces - 1 row ## 3 Currier 1970 Quad 1 merces - 1 row ## 4 Mather 1970 River East 1 merces - 1 row ## 4 Mather 1970 River East 1 merces - 1 row ## 4 Mather 1970 River East 1 merces - 1 row cetor ict+</pre>		#	#	4 004	Ecor	L Ma	ather	9					
<pre>## 6 006 Stat Winthrop 7 ## 7 007 IB Pfoho 8 houses ## 1 Dunster 1930 River East 0 methods 7 1000 ## 2 Winthrop 1931 River West 0 methods 7 1000 ## 3 Currier 1970 Quad 1 methods 7 1000 ## 4 Mather 1970 River East 1 methods 7 1000 cereous cereos 1000 cer</pre>	right	#	#	5 005	Psych	L I	Pfoho	6					
<pre>## 7 007 IB Proho 8 houses ## name built area ## 1 Dunster 1930 River East 0 merces 7 1 read ## 2 Winthrop 1931 River West 3 merces 7 1 read ## 4 Mather 1970 River East 1 merce 7 1 read ## 4 Mather 1970 River East 1 merce 7 1 read for any 1 merce 7 1 read for a sect 1 merce 7 1 merce 7 1 read for a sect 1 merce 7 merce 7</pre>		#	#	6 006	Stat	Wint	throp	7					
<pre>houses houses ## name built area ## 1 Dunster 1930 River East 0 metdes - 1000 ## 2 Winthrop 1931 River West 3 metdes - 3005 ## 3 Currier 1970 River East 1 metdes - 1000 ## 4 Mather 1970 River East 1 metdes - 1000 for metch but kept left_join(houses, students, from join_by("name" == "house")) ## name built area id conc sleep ## 1 Dunster 1930 River East <na> NA T ## 2 Winthrop 1931 River West 001 CPB 7 S ## 3 Winthrop 1931 River West 003 Stat 8 metchs ## 4 Winthrop 1931 River West 006 Stat 7 ## 6 Mather 1970 River East 004 Econ 9 </na></pre>		## 7 007 IB Pfoho 8											
<pre>1 ## name built area 1 eft ## 1 Dunster 1930 River East 0 metrics - 3 metrics ## 3 Currier 1970 Quad metrics - 3 metrics ## 4 Mather 1970 River East 1 metrics - 1 metrics ## 4 Mather 1970 River East 1 metrics for action of the second seco</pre>		houses											
<pre> 1 cf.t ## 1 Dunster 1930 River East 0 methods 7 1933 ## 3 Currier 1970 Quad 1 methods 7 1930 ## 4 Mather 1970 River East 1 methods 7 1930 ## 4 Mather 1970 River East 1 methods 7 1930 ## 1 Dunster 1930 River East 1 methods 7 1930 ## 1 Dunster 1930 River East 1 methods 7 1930 ## 1 Dunster 1930 River East 1 methods 7 1930 ## 1 Dunster 1930 River East 1 methods 7 1930 ## 1 Dunster 1930 River East 1 methods 7 1930 ## 2 Winthrop 1931 River West 001 CPB 7 5 ## 3 Winthrop 1931 River West 003 Stat 8 methods 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9 ## 6 Mather 1970 River East 004 Econ 9 </pre>	`	#	#		name	built	t a n.	area	1				
<pre>## 3 Currier 1970 Quad Quad marches - jrew ## 4 Mather 1970 River East marches - jrew Gereve / cftord bot kept left_join(houses, students, prom join_by("name" == "house")) "ident j</pre>	1eft	#	#	1 Dun 2 Wint	brop	1930	J Rive	r East	. 0 m	tones	- 7.00		
<pre>## 4 Mather 1970 River East Imarcase 7 1000 6 0000 ceter /cetoon() 6 000 ceter /cet+_oon() 6 000 ceter /ceter /cet+_oon() 6 000 ceter /ceter /cet</pre>	1	#	#	3 Cur	rier	1970		Quad	1 1 -	whenes	-> I row		
<pre>6 rows cerc /cetord) no match, but kapt left_join(houses, students, room join_by("name" == "house")) "'cert" ## name built area id conc sleep ## 1 Dunster 1930 River East <na> <na> NA T ## 2 Winthrop 1931 River West 001 CPB 7 5 ## 3 Winthrop 1931 River West 003 Stat 8 metures ## 4 Winthrop 1931 River West 006 Stat 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9</na></na></pre>		#	#	4 Ma	ther	1970) Rive	r East	lm	arches	-> Irow		
<pre>no match; but kept left_join(houses, students, from join_by("name" == "house")) "" "" "" "" "" "" "" "" "" "" "" ""</pre>													
<pre>no match; but kept left_join(houses, students,</pre>	6												
<pre>no match; bot kept left_join(houses, students,</pre>													
<pre>bot kept left_join(houses, students,</pre>											5		
<pre>kept left_join(houses, students,</pre>	no match										5		
<pre>ideal_join_by("name" == "house")) ## name built area id conc sleep ## 1 Dunster 1930 River East <na> <na> NA ## 2 Winthrop 1931 River West 001 CPB 7 s ## 3 Winthrop 1931 River West 003 Stat 8 ## 4 Winthrop 1931 River West 006 Stat 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9</na></na></pre>	no match	1									5		
indicationjoin_by(nameindication##namebuiltareaid conc sleep##1Dunster1930RiverEast <na><na>T##2Winthrop1931RiverWest001S##3Winthrop1931RiverWest003StatS##4Winthrop1931RiverWest006Stat7##5Currier1970Quad002HDRB8##6Mather1970RiverEast004Econ9</na></na>	no match but kept	۲ let	ft	ioin(house	s. st	udent	5.			5		
##name builtareaid conc sleep##1Dunster1930River East <na>NAT##2Winthrop1931River West001CPB75##3Winthrop1931River West003Stat8##4Winthrop1931River West006Stat7##5Currier1970Quad002HDRB8##6Mather1970River East004Econ9</na>	no match but kept prom	י lei	[t_	join(house	s, st	udent	s,					
★ ## 1 Dunster 1930 River East <na> <na> NA T ## 2 Winthrop 1931 River West 001 CPB 7 5 ## 3 Winthrop 1931 River West 003 Stat 8 merces ## 4 Winthrop 1931 River West 006 Stat 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9</na></na>	no match but kept prom original	lei	ft_	join(house join_	s, st by("r	udent:	s, == "ho	ouse"))			
## 1 Dunster 1930 River East NA T ## 2 Winthrop 1931 River West 001 CPB 7 5 ## 3 Winthrop 1931 River West 003 Stat 8 merces ## 4 Winthrop 1931 River West 006 Stat 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9	no match but kept prom original "left" d	led	ft_ sct	join()	house join_ ame_b	s, st by("r	udent: name" :	s, == "ho	ouse") id) conc	sleep		
## 2 Winthrop 1931 River West 001 CPB 7 5 ## 3 Winthrop 1931 River West 003 Stat 8 merces ## 4 Winthrop 1931 River West 006 Stat 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9	no match but kept original "ictt" o	le:	ft_ sct	join() n	house join_ ame b	s, st by("r uilt	udent: name" :	s, == "ho area	id) conc	sleep		
5 ## 3 Winthrop 1931 River West 003 Stat 8 mercines ## 4 Winthrop 1931 River West 006 Stat 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9	no match but kept original "ict" o	le:	ft_ set	join() n Duns	house join_ ame b ter	s, st by("r uilt 1930	River	s, == "ho area East	ouse") id <na></na>) conc <na></na>	sleep		
## 4 Winthrop 1931 River West 006 Stat 7 ## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9	no match but kept original "left" o	le: ## ## ##	ft_ sct 1 2	join () n Duns Winth	house join_ ame b ter rop	s, st by("r uilt 1930 1931	River	s, == "ho area East West	ouse") id <na> 001</na>) conc <na> CPB</na>	sleep NA 7		
## 5 Currier 1970 Quad 002 HDRB 8 ## 6 Mather 1970 River East 004 Econ 9	no match but kept original "left" T	le: ## ## ##	ft_ sct 1 2 3	join() n Duns Winth Winth	house join_ ame b ter rop rop	s, st by("r uilt 1930 1931 1931	River River River	s, == "ho area East West West	ouse") id <na> 001 003</na>) conc <na> CPB Stat</na>	sleep NA 7 8		
## 6 Mather 1970 River East 004 Econ 9	no match but kept viett J	le1 ## ## ## ##	ft_ sct 1 2 3 4	join() n Duns Winth Winth Winth	house join_ ame b ter rop rop rop	s, st by("r 1930 1931 1931 1931	River River River River River	s, == "ho area East West West West	id <na> 001 003 006</na>) conc <na> CPB Stat Stat</na>	sleep NA 7 8 7		
	no match but kept original "Ict" J	led ## ## ## ## ##	ft_ sct 1 2 3 4 5	join() n Duns Winth Winth Winth Curr	house join_ ame b ter rop rop rop ier	s, st by("r 1930 1931 1931 1931 1931	River River River River River	s, area East West West West Quad	id <na> 001 003 006 002</na>	conc <na> CPB Stat Stat HDRB</na>	sleep NA 7 8 7 8		
	no mater but kept droiginal "ict" 5 marches	le: ## ## ## ## ##	ft_ 1 2 3 4 5 6	join () n Duns Winth Winth Winth Curr Mat	house join_ ame b ter rop rop rop ier her	s, st by("r 1930 1931 1931 1931 1970 1970	River River River River River	s, area East West West Quad East	id <na> 001 003 006 002 004</na>	conc <na> CPB Stat Stat HDRB Econ</na>	sleep NA 7 8 7 8 9		

<mark>Inner Join</mark>

- inner_join(houses, students, join_by("name" == "house"))
- Combine 2 datasets via key, keeping only matching observations between BOTH datasets (most constrained)

	5	tu	dents							
	#	#	id 1 001	CPB	h Wint	ouse s brop	leep 7			
	#	#	2 002	HDRB	Cur	rier	8			
	#	##	3 003	Stat	Wint: Ma	hrop ther	8			
	#	#	5 005	Psych	P	foho	6			
	#	##	6 006 7 007	Stat TB	Wint	hrop fobo	7			
	h	ou	ses	10		10110	U			
	#	#		name	built		area			
	#	⊧# ⊧#	1 Dur 2 Wint	nster throp	1930 1931	River River	East West			
	#	#	3 Cui	rrier	1970		Quad			
	#	#	4 Ma	ather	1970	River	East			
	inr		. ioin	(house		tudent	ts			
				d a d a	b(I	"			
				Join	_by("	name	== "1	iouse	•··))	
	##		n	name b	uilt		area	id	conc	sleep
т	##	1	Winth	rop	1931	River	West	001	CPB	7
1	##	2	Winth	iron	1931	River	West	003	Stat	8
	ππ	2	W TH 01	rr ob	1001	TUTVET	WCBC	000	Duau	0
5		0			1001	D ·		000	a .	-
5 the	##	3	Winth	rop	1931	River	West	006	Stat	7
5 tohes	>## ##	3 4	Winth Curr	rop ier	1931 1970	River	West Quad	006 002	Stat HDRB	7 8
5 tones	>## ##	3 4 5	Winth Curr Mat	rop ier ther	1931 1970 1970	River	West Quad East	006 002 004	Stat HDRB Econ	7 8 9
5 Hones	>## ## ##	3 4 5	Winth Curr Mat	rop Tier Ther	1931 1970 1970	River River	West Quad East	006 002 004	Stat HDRB Econ	7 8 9

ma

Full Join

- full_join(houses, students, join_by("name" == "house"))
- Combine 2 datasets via key, keeping all observations between BOTH datasets and putting N/A if an observation didn't have corresponding value for a variable (most expansive)

	stu	de	ents								
	##		id co	onc	h	ouse	s]	leep			
	##	1	001 0	CРВ	Winth	nrop		7			
	##	2	002 HI	DRB	Curr	rier		8			
	##	з	003 St	at	Winth	nrop		8			
	##	4	004 Ec	on	Mat	ther		9			
	##	5	005 Psy	rch	P	foho		6			
	##	7	006 50	тр	WINT	fobo		6			
	hou	' ISe	s	тв	F .	1 0110		0			
	##		nan	ne 1	built			area			
	##	1	Dunste	er	1930	Rive	r	East			
	##	2	Winthro	\mathbf{p}	1931	Rive	r	West			
	##	з	Currie	\mathbf{r}	1970			Quad			
	##	4	Mathe	er	1970	Rive	r	East			
	fu	11	ioin(ho	nise	es. st	uden	ts	-			
			-J	in	by ("r	ame"	_	, = "ho	1190"		
			J	· 111_	_by(I	lane	_	- 110	use ,	,	
	##		nar	ne 1	built			area	id	conc	sleep
1	##	1	Dunste	er	1930	Rive	r	East	<na></na>	<NA $>$	NA
/ T	##	2	Winthro	pp	1931	Rive	r	West	001	CPB	7
	##	з	Winthro	pp	1931	Rive	r	West	003	Stat	8
matche	5##	4	Winthro	pp	1931	Rive	r	West	006	Stat	7
1	##	5	Currie	\mathbf{r}	1970			Quad	002	HDRB	8
	##	6	Mathe	er	1970	Rive	r	East	004	Econ	9
10	##	7	Pfol	10	NA			<na></na>	005	Psych	6
	##	8	Pfoł	10	NA			<na></na>	007	IB	8
24			thes hu	+ =+	-11 - 22	- 2					

Practice: Answer These Questions

- LH dataset has 12 houses; RH dataset has 30 students (but not every student has to be in a house!)
- For left_join(), where houses is the LH dataset, what is the min # of rows possible?
- For inner_join(), what is the min/max # of rows possible?
- For full_join(), what is the min/max # of rows possible?

Solution: Left Join

For left_join(), there must be at least 12 rows because each of the 12 rows in houses has to be represented, even if there are no matches for students (for example, if all 30 students are first-years)



Solution: Inner Join

- For inner_join(), there can be o rows if there are no matches between students and houses, and there can be up to 30 rows if all students are matched to a house
- It may seem like there'd be a maximum of 12 rows, but if all 30 students had Winthrop, there'd be 30 rows (see Slide 17)



Solution: Full Join

For full_join(), there can be 30 rows in the event that all students are matched to a house, and there can be up to 42 rows if there are no matches between students and houses



Removing Missing Values

https://drive.google.com/file/d/1ZMcx2lXfGUTBu aEdbYsHyA6O8doTKLEa/view?usp=sharing

Important Code for Data Wrangling

https://drive.google.com/file/d/1z56My5Te6hX_J 6iIXOAMooBIx5mxZopa/view?usp=sharing



- <u>%>%</u>: Takes dataset and "pipes" it as the first argument in the next line
 - The first argument of most wrangling verbs is a dataset
 - This is read as "and then" when reading code aloud
- 'Command' + 'Shift' + 'M'

Pipe: These Are Equivalent Statements

mythbusters %>%

summarize(count = n())

summarize(mythbusters, count = n())

Practice 1: What Does This Code Do?

women_in_stem <- people %>%

filter(gender == "Female", jobtitle == "Software Engineer" | jobtitle == "Mathematician")

Practice 2: What Does This Code Do?

students <- students %>%

mutate(seniority_new = case_when(

seniority <= 2 ~ "junior",

seniority == 3 ~ "mid",

seniority >= 4 ~ "senior"))

Practice 3: What Does This Code Do?

people %>%

drop_na(pay) %>%

filter(gender == "Female", jobtitle == "Financial Analyst") %>%

slice_max(pay, n = 10) %>%

select(pay, education, name)

Questions?

P-Set 2

Have a great rest of your week!