STAT 102: 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

Join Our Slack: #section-d001



Content Review: Week 2

Grammar of Graphics

<u>https://drive.google.com/file/d/1y13AW7qNlvMn</u> <u>HGtNv1FspjUb4Vfr9Ohn/view?usp=drive_link</u>

Choosing the Right Graph

<u>https://drive.google.com/file/d/1GlfYFuUYMPxM</u> <u>gnvBtojxzQn6MGh5yl-H/view?usp=drive_link</u>



- 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	🖸 🍝 US 🔿 4
🔍 🔍 💼 📥 STAT 10	0 Teachii X 🚦 STAT 100 Sectio: X 📃 STAT 100 Sectio: X 💧 STAT 100 - Rem: X 🎲 Course Roster: S X	💥 All Content - STA X 🛛 🧱 Prol
$\leftarrow \rightarrow \mathbf{C}$	C A == https://posit.cloud/spaces/534915/content/8756398	\$
🕅 Notion 🔲 Calendar 🛛 Main Mai	I 🔰 School Mail 🚍 Main Docs 😑 School Docs 💧 Main Drive 💧 School Drive 👳 Harvard 🔅 Canvas 🛐 Sy	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	
A Your Workspace	🔍 🔹 🚽 🚰 🔚 🚔 🖌 🖉 Co to file/function	
() Tour Workspace	🗢 pset02_stat100_fall2024.Rmd × 📃 colleges × 📃 colleges_moderate_removal × 📃 colleges_) >>> 🗇	Environment History Conne
G STAT 100 Fall 2024 Harvard University	_ (==)	💣 🔒 📑 Import Dataset 🔹 🌾
	Source Visual 🖻 Outline	R 👻 🐴 Global Environment 👻
New Space Learn Guide What's New	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 ("mr_kd5_pql'). Create a plot that shows the association between 'mr_kd5_pql' and average annual cost of attendance; customize the color and transparency of the 'geom'. Describe what you see.] 338 339 * ```{r} 340 341 ggplot(data = colleges, mapping = aes(y = sticker_price_2013, x = sat_avg_2013, x = sat	Data • colleges 1285 • colleges_aggressi. 279 c • colleges_light_re. 1285 • colleges_moderate. 1284 • Colleges_moderate. 1284
Recipes Cheatsheets	<pre>color = tier_nome() + 32 geom_sonot(alpha = 0.5) + 33 geom_smoot(inmethod = "lm", se = FALSE) + 34 lobs(s = "Average SAT in 2013", y = "Sticker Price in 2013", color = "Tier Name", title = "Colleges in America") 345 346 347 347 346</pre>	Rev Folder New Folder New Folder New Blank Fil Coud > project A Name Shistory data
Help		B project Roroj
🌵 Current System Status <		 project.cproj pset02_stat100_fall202 pset02_stat100_fall202
Info \$ Plans & Pricing	Ø [38;5:232m ³ geom_smooth0 ³ using formula = 'y ~ x'[39m ▲ Warning: [38;5:232menoved 341 rows containing non-finite outside the scale range ('stat_smooth0 ³);33 ^m Warning: [38;5:232mRemoved 341 rows containing missing values or values outside the scale range ('geom_point0 ³);39 ^m	
Terms and Conditions	Collogos in America	
	537.349 🖬 Problem 4 V R Markdown C	
	Console	

<mark>Clean Code</mark>

🔹 🌢 STAT 100 Teachi X 🛗 STAT 100 Sectio X 🔤 STAT 100 Sectio X 💩 STAT 100 - Rem: X 🎲 Course Roster: S X 🐹 All Content - STA	× 🔀 Prot
← → C () A ≈ https://posit.cloud/spaces/534915/content/8756398 ☆	
🗓 Notion 🔟 Calendar 🤘 Main Mail 🔰 School Mail 🚍 Main Docs 🚍 School Docs 🝐 Main Drive 🝐 School Drive 🎅 Harvard 🎲 Canvas 🛐 Symbolab 📊 Gradescop	e 🐹 Posit C
STAT 100 Fall 2024 / Problem Set 2	
Spaces File Edit Code View Plots Session Build Debug Profile Tools Help	
Vour Workspace	
pset02_stat100_fall2024.Rmd × colleges × colleges_moderate_removal × colleges_> Finvironment Hist	ory Conne
(2) STAT 100 Fall 2024 → □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □	Dataset 👻 🗌
Source Visual R + 💁 Global Envi	ronment +
T New Space 337 a) Let's examine one measure or mobility force: the percentage of students with porents in the bottom income quintile (Cimr_kq5,pq1). Create a plot that shows the association between mr_kq5,pq1 and ocleges.gape or allower the students with geom. Bescribe what you see. Data Data Learn isgem. Bescribe what you see. isgem. Colleges.inde isgem. Students with e students with out on and transparency of the geom. Bescribe what you see. isgem. Students with e students withe students with e students with	1285 ssi 279 c _re 1285 ate 1284 kages Help New Blank Fil
Q Cheatsheets 346 lobs(x = "Average SAT in 2013", 347 ************************************	oj t100_fall202
Posit Community	t100_fall202
Into	
\$ Plans & Pricing ▲ Warning: [38,5;232m 'geom _smooth') using formula = y ~ x'[39m ▲ Warning: [38,5;232mRemoved 341 rows containing non-finite outside the scale range ('stat_smooth') 'J39m	
Terms and Conditions Maximum (29:6:323mBomound 241 sour containing missing values or values of the strength of	

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

	st	ud	ents							
	##	E .	id	con	c	house	sleep			
	##	± 1	001	CP	B Win	throp	7			
1	##	2	002	HDR	B Cu	rrier	8			
	##	: 3	003	Sta	t Win	throp	8			
right	##	. 4	004	Psvc	h M	Pfobo	9			
•0	##	• 6	006	Sta	t Win	throp	7			
	##	• 7	007	I	в	Pfoho	8			
	hc	us	es							
-	##	E .		name	buil	t	area	ı		
left	##	± 1	Dur	nster	193	O Rive	r East	0~	tches	-> I row
	##	2	Wint	throp	193	1 Rive	r West	3 m	arches	-> 3 rous
	##	: 3	Cui	rrier	197		Quad		when	- 100
	###	- 4	1414	ather	197	O KIVE	r cast			
									6	Cows
								~F	to lef	+-Join()
no match	L.									
no match but	ı									
no match but kept]	.ef	t_j	oin(hous	es, st	tudent	s,			
no match but kept] from	.ef	t_j	oin(hous	es, st bv("1	udent:	s, == "ho	ouse")		
no match but kept 1 prom original	.ef	t_j	oin(hous	es, st _by("1	tudent; name"	s, == "ho	ouse"))	
no match but kept 1 prom original "left" da	.ef +///	t_j ct	oin(hous join	es, st _by("1 built	tudent: name"	s, == "ho area	ouse") id) conc	sleep
no match but kept 1 prom original "Ist" da	.ef +7×5 #	t_j ct	oin(n Duns	hous join ame	es, st _by("r built 1930	tudent: name"	s, == "ho area Fast	ouse") id) conc	sleep
no match but kept 1 crom original "isti" sa	.ef ***	t_j ct 1	oin(n Duns	hous join ame ter	es, st _by("1 built 1930	River	s, == "ho area East	ouse") id <na></na>) conc <na></na>	sleep NA
no match byt croint vist *	.ef # # #	t_j et 1 2 h	oin(n Duns Jinth	hous join ame ster arop	es, s ¹ _by("1 built 1930 1931	River	s, == "ho area East West	ouse") id <na> 001</na>) conc <na> CPB</na>	sleep NA 7
no match bypt poriginal view of ###	.ef ## ## ##	t_j et 1 2 h 3 h	oin(n Duns Jinth Jinth	hous join ame ter arop arop	es, st _by("1 built 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
no watch but kopt biginel viet big t t t t t t t t t t t t t t t t t t t	.ef # # # #	t_j et 1 2 h 3 h 4 h	oin(n Duns Vinth Vinth	hous join ame ter arop arop	es, st _by("1 1930 1931 1931 1931	River River River River River	s, == "ho area East West West West	id <na> 001 003 006</na>) <na> CPB Stat Stat</na>	sleep NA 7 8 7
no match byt com original big incl big incl big incl big incl big incl big incl big incl big incl big incl big incl big t incl incl big t i i i i i in incl big t i i i i i i i i i i i i i i i i i i	ef # # # # #	t_j et 1 2 h 3 h 4 h	n Duns Jinth Jinth	house join ame ster arop arop	es, st _by("r built 1930 1931 1931 1931	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 byt cropt original origin original original original original orig	ef ## ## ## ##	t_j et 1 2 h 3 h 4 h	oin(Duns Jinth Jinth Curr	house join ame ter arop arop arop	es, st _by("1 built 1930 1931 1931 1931 1970	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 match byopt byo	.ef ## ## ## ## ##	t_j 2 k 3 k 4 k 5	oin(Duns Jinth Jinth Jinth Curr Mat	hous join ame ter trop trop ter her	es, st _by("1 built 1930 1931 1931 1931 1970 1970	River River River River River River	s, == "ho East West 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	conc	h	ouse s	leep			
	#	#	2 002	HDRB	Cur	rier	8			
	#	#	3 003	Stat	Wint	hrop	8			
	#	#	4 004 5 005	Psych	Ma P	ther foho	6			
	#	#	6 006	Stat	Wint	hrop	7			
	# h	*#	7 007	IB	P	foho	8			
	#	#		name	built		area			
	#	#	1 Du	nster	1930	River	East			
	#	#	2 Win 3 Cu	throp rrier	1931	River	West Quad			
	#	#	4 Ma	ather	1970	River	East			
				()						
	ınr	ler		(house	es, s	tudent	ts,			
				join	_by("	name"	== "}	iouse	e"))	
	##		r	name b	uilt		area	id	conc	sleep
т	##	1	Winth	rop	1931	River	West	001	CPB	7
1	##	2	Winth	rop	1931	River	West	003	Stat	8
-	##	3	Winth	iron	1931	River	West	006	Stat	7
5	###					TOTACT	1000	000	Sout	
S Home	5 ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	^	~ ~ ~		1070		0 1	000	IIDDD	0
S tches 1	>##	4	Curi	rier	1970		Quad	002	HDRB	8
S Hone	>## ##	4 5	Curr Mat	ier her	1970 1970	River	Quad East	002 004	HDRB Econ	8 9

mu



- 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									
	000	lde	ents							
	##	1	id	CODE	h	ouse s	sleep			
	##	2	001	HDBB	Curr	rier	8			
	##	З	003	Stat	Wintl	hrop	8			
	##	4	004	Econ	Ma	ther	9			
	##	5	005	Psych	P:	foho	6			
	##	6	006	Stat	Wint	hrop	7			
	##	'	007	16	P.	10110	0			
	##	ise	.5	namo	built		2702			
	##	1	Dur	name	1930	River	East			
	##	2	Wint	hrop	1931	River	West			
	##	з	Cur	rier	1970		Quad			
	##	4	Ma	ather	1970	River	East			
	fu	11.	_joi	n(hous join	es, st _by("r	udent	s, == "ho	ouse")))	
	fu:	11.	_joi	n(hous join name	es, st _by("r built	udent name"	s, == "ho area	ouse") id)) conc	sleep
<u></u>	fu: ## < ##	11.	_join	n(hous join name nster	es, st _by("r built 1930	udent name" River	s, == "ho area East	ouse") id <na></na>)) conc <na></na>	sleep NA
	fu: ## ## ##	11 1 2	_join Dur Win	n(hous join name nster throp	es, st _by("r built 1930 1931	udent name" River River	s, == "ho area East West	id <na> 001</na>	conc <na> CPB</na>	sleep NA 7
	fu: ## ## ##	11 1 2 3	_join Dur Win [.] Win [.]	n(hous join name nster throp throp	es, st _by("r built 1930 1931 1931	River River River River	s, == "ho area · East · West · West	ouse") id <na> 001 003</na>)) conc <na> CPB Stat</na>	sleep NA 7 8
	fu: ## ## ##	11 1 2 3 4	_join Dur Win [.] Win [.] Win [.]	n(hous join name nster throp throp throp	es, st _by("r built 1930 1931 1931 1931	River River River River River	s, == "ho East West West West	id <na> 001 003 006</na>)) conc <na> CPB Stat Stat</na>	sleep NA 7 8 7
	fu: ## ## ## \$## ##	11 2 3 4 5	join Dur Win Win Cur	n(hous join name nster throp throp throp rrier	es, st _by("r built 1930 1931 1931 1931 1931	River River River River River	s, == "ho 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
s antrine	fu: ## ## ## \$## ##	11. 1 2 3 4 5 6	_join Dun Win Win Win Cun Ma	n(hous join name nster throp throp throp rrier ather	es, st _by("r 1930 1931 1931 1931 1970 1970	River River River River River River	s, == "ho 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 7 8 9
	fu: ## ## ## \$## ##	11 2 3 4 5 6 7	_join Dun Win Win Cun Ma	n (hous join name nster throp throp throp rrier ather Pfoho	es, st _by("r built 1930 1931 1931 1931 1970 1970 NA	River River River River River River	s, == "ho area East West West Quad East <na></na>	id <na> 001 003 006 002 004 005</na>	conc <na> CPB Stat Stat HDRB Econ Psych</na>	sleep NA 7 8 7 8 9 6
	fu: ## ## ## ## ##	11. 12 34 56 78	_join Dur Win Win Cur Ma	n (hous join name nster throp throp throp rrier ather Pfoho	es, st _by("r built 1930 1931 1931 1931 1970 1970 NA	River River River River River River	s, == "ho East West West Quad East <na></na>	id <na> 001 003 006 002 004 005 007</na>	conc <na> CPB Stat Stat HDRB Econ Psych TB</na>	sleep NA 7 8 7 8 9 6 8

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() and 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

<u>https://drive.google.com/file/d/1poOyUn9yBChK</u> <u>crBxGZDiTleOfKIqDWYm/view?usp=drive_link</u>

Important Code for Data Wrangling

<u>https://drive.google.com/file/d/1Mr7SHkdQo6N</u> <u>G8k4BfHrr5HSg1CEJrrL/view?usp=drive_link</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")



- We define a new dataset "women_in_stem" assigned from the dataset "people"...
- And then we filter for the gender of "female" and the job title of "software engineer" or "mathematician"

Practice 2: What Does This Code Do?

students_new <- students %>%

mutate(seniority_new = case_when(

seniority <= 2 ~ "junior",</pre>

seniority == 3 ~ "mid",

seniority >= 4 ~ "senior"))



- We **assign** the **dataset** "students" to itself (so that it gets updated)...
- And then we mutate a new variable "seniority_new", which is...
 - Equal to "junior" when the variable "seniority" is less than/equal to 2
 - Equal to "mid" when the variable "seniority" is equal to 3
 - Equal to "senior" when the variable "seniority" is greater than/equal to 3

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)



- We take the **dataset** "people"...
- And then we drop all observations that have NA for the variable "pay"...
- And then we filter for the gender of "female" and the job title of "financial analyst"
- And then we slice for the observations with the top 10 maximum values for the variable "pay"
- And then we select (to display) the variables "pay," "education," and "name"

Questions?

P-Set 2

Have a great rest of your week!