

STAT 100: Week 3

Ricky's Section

Introductions and Attendance

Introduction: Name

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

Important Reminders

Anonymous Feedback

https://docs.google.com/forms/d/e/1FAIpQLSfKv_FGvs0oqm-IvtxKx3Vf6bBzSJE2jamK1gklAzL6NkXE8w/viewform

Join our Slack: #section-d003

<https://join.slack.com/share/enQtNzY3NDUxNTQxMzkwOS1lNzVhZWQxN2ZmYWMyYThlMzIyYmM4ZmNiMTJmOWViZGE2MzQ5NWVkMTM2YWY4MWRhMWZiZDAxOWZlZDYxYmYx>

Content Review: Week 2

Grammar of Graphics

https://drive.google.com/file/d/1nAVN8_qi1vWCwllXjgat7MukSv5Wgpqm/view?usp=sharing

Choosing the Right Graph

<https://drive.google.com/file/d/1T1rovHS3l1rapzRH6fGNclS2fn6lu7gs/view?usp=sharing>

Coding

- 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 (+)

12 1 United Talmudical Seminary 21200

Problem 4

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_kq5_pq1). Create a plot that shows the association between mr_kq5_pq1 and average annual cost of attendance; customize the color and transparency of the geom. Describe what you see.

```
ggplot(data = colleges, mapping = aes(y = sticker_price_2013, x = sat_avg_2013, color =
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", titl
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 341 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Warning: Removed 341 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

Problem Set 2 - 1 X Problem Set 2 - 1 X + -

RAM Settings Profile Ricky Truong

R 4.4.1 Tutorial 474 MiB List

Table with 2 columns: Size, Modified. Rows include '0 B Aug 15, 2024, 5:46 PM', '205 B Sep 18, 2024, 12:46 PM', '408.3 KB Sep 18, 2024, 12:49 PM', '23.5 KB Sep 18, 2024, 12:48 PM'.

Messy Code

```
337 a) Let's examine one measure of mobility rate: the percentage of students with
338 parents in the bottom income quintile who ended up in the top income quintile
339 ('mr_kq5_pq1'). Create a plot that shows the association between 'mr_kq5_pq1' and
340 average annual cost of attendance; customize the color and transparency of the
341 'geom'. Describe what you see.
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

Warning: [38;5;232mRemoved 341 rows containing non-finite outside the scale range ('stat_smooth0').]39m

Warning: [38;5;232mRemoved 341 rows containing missing values or values outside the scale range ('geom_point0').]39m

Clean Code

```
337 a) Let's examine one measure of mobility rate: the percentage of students with
338 parents in the bottom income quintile who ended up in the top income quintile
339 ('mr_kq5_pq1'). Create a plot that shows the association between 'mr_kq5_pq1' and
340 average annual cost of attendance; customize the color and transparency of the
341 'geom'. Describe what you see.
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

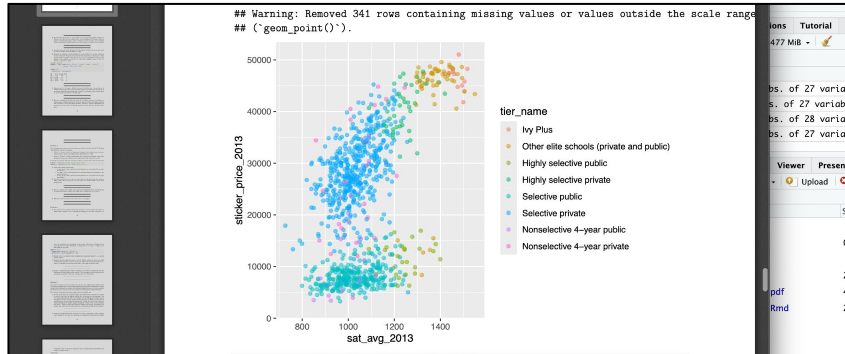
Warning: [38;5;232mRemoved 341 rows containing non-finite outside the scale range ('stat_smooth0').]39m

Warning: [38;5;232mRemoved 341 rows containing missing values or values outside the scale range ('geom_point0').]39m

Reading Code in English

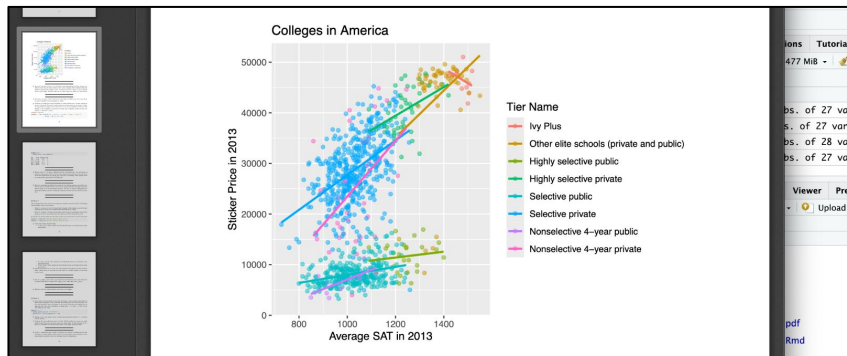
In R: `ggplot(data = colleges, mapping = aes(y = sticker_price_2013, x = sat_avg_2013, color = tier_name)) + geom_point(alpha = 0.5)`

In English: 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

In R: `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")`



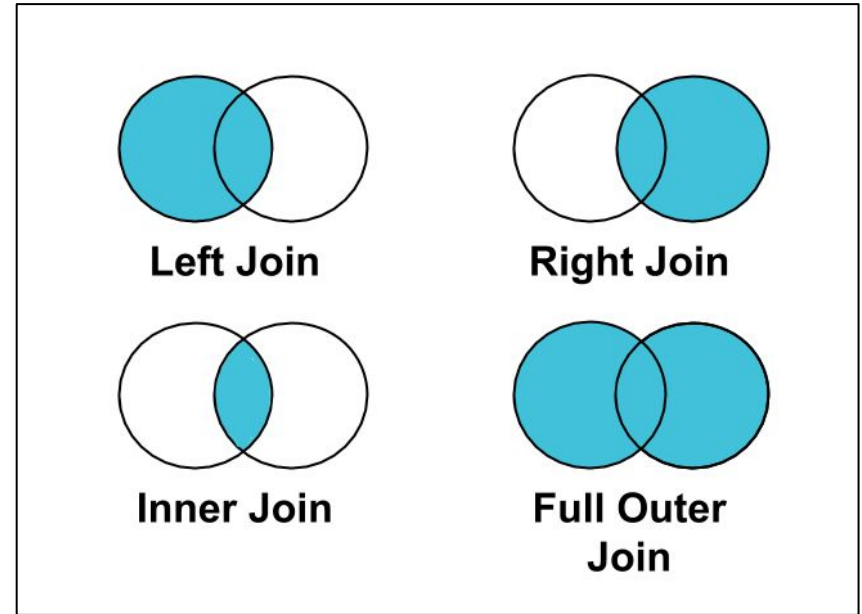
In English: Create a **plot** using the “colleges” **dataset**. Map to the **aesthetic** of **y-direction** the **variable** of sticker price, map to the **aesthetic** of **x-direction** the **variable** of average SAT, and map to the **aesthetic** of **color** the **variable** tier name. Use **points** 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

Data Joins

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



Left Join

- `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

```
students
##   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
```

right

```
houses
##   name built area
## 1 Dunster 1930 River East
## 2 Winthrop 1931 River West
## 3 Currier 1970 Quad
## 4 Mather 1970 River East
```

left

0 matches → 1 row
3 matches → 3 rows
1 matches → 1 row
1 matches → 1 row

6 rows after left_join()

```
no match, but kept from original "left" dataset
left_join(houses, students,
           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
## 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
```

matches

Inner Join

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

```
students
##   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
##   name built   area
## 1 Dunster 1930 River East
## 2 Winthrop 1931 River West
## 3 Currier 1970 Quad
## 4 Mather 1970 River East
```

```
inner_join(houses, students,
           join_by("name" == "house"))
```

```
##   name built   area id conc sleep
## 1 Winthrop 1931 River West 001 CPB 7
## 2 Winthrop 1931 River West 003 Stat 8
## 3 Winthrop 1931 River West 006 Stat 7
## 4 Currier 1970 Quad 002 HDRB 8
## 5 Mather 1970 River East 004 Econ 9
```

T
S
matches
L

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)

```
students
##   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
##   name built   area
## 1 Dunster 1930 River East
## 2 Winthrop 1931 River West
## 3 Currier 1970   Quad
## 4 Mather 1970 River East
```

```
full_join(houses, students,
          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
## 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
## 7 Pfoho   NA   <NA> 005  Psych    6
## 8 Pfoho   NA   <NA> 007   IB     8
```

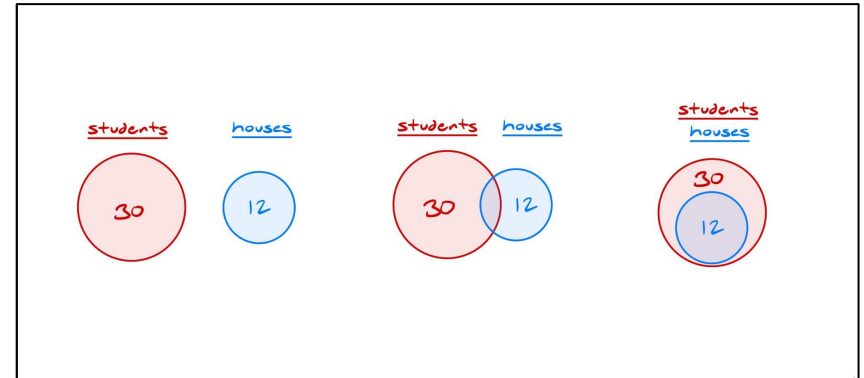


Practice: Answer These Questions

- 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()`, what is the min/max # of rows possible?
- For `full_join()`, what is the min/max # of rows possible?

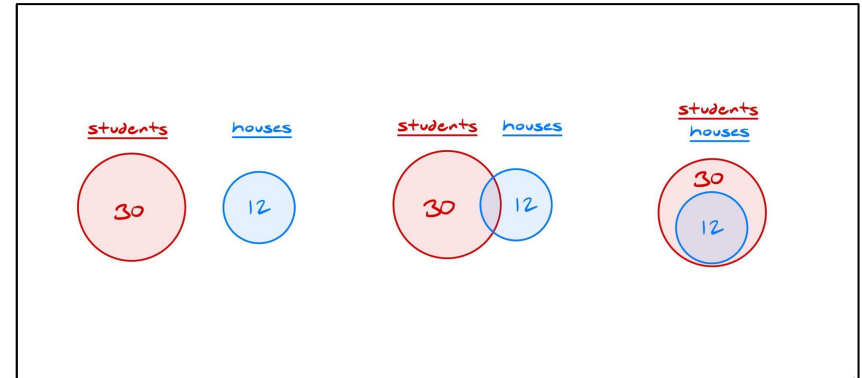
Solution: Left Join

- 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)



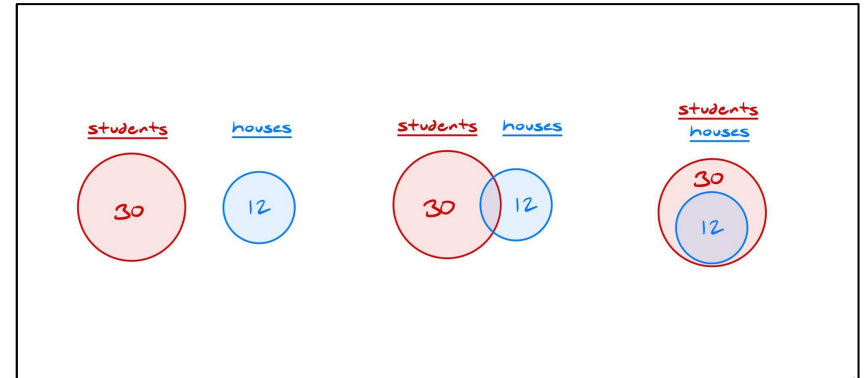
Solution: Inner Join

- For `inner_join()`, there can be 0 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_I6iIXOAMooBIx5mxZopa/view?usp=sharing

Pipe

- %>%: 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!