Data manipulation

This is the NBA games data that we worked with previously in class and on assignments. The first 6 rows are shown below.

```
g = readRDS('data/games.rds')
gg = g %>%
  filter(lg == 'nba',
         season %in% 2022,
         season.type == 'reg') %>%
  select(date, away, home,
         ascore, hscore, season, gid, lg) %>%
  mutate(gid = as.numeric(gid) - 22100000) %>%
  arrange(gid)
head(gg)
```

```
        date away home ascore hscore season gid  lg
1 2021-10-19  BKN  MIL    104    127   2022   1 nba
2 2021-10-19  GSW  LAL    121    114   2022   2 nba
3 2021-10-20  IND  CHA    122    123   2022   3 nba
4 2021-10-20  CHI  DET     94     88   2022   4 nba
5 2021-10-20  BOS  NYK    134    138   2022   5 nba
6 2021-10-20  WAS  TOR     98     83   2022   6 nba
```

Suppose we run the the following code.

```
a = gg %>% select(gid, away, ascore); colnames(a) = c('gid', 'team', 'score')
h = gg %>% select(gid, home, hscore); colnames(h) = c('gid', 'team', 'score')

d = rbind(a,h) %>%
  arrange(gid)
```

Write the first four rows of `d` below.

a

gid	away	~~bcore~~	ascore
1	BRN	104	
2	GSW	121	

h

1	MIL	127
2	LAL	114

d

1	BRN	104
1	MIL	137
2	GSW	121
2	LAL	114

Commenting code

Below are the first three and last three rows of a US Census Age and Sex data data set.

```
# A tibble: 3 x 4
   year mf     age.group     pop
  <dbl> <chr>  <fct>       <dbl>
1  2010 male   0 to 4    2018474
2  2011 male   0 to 4    2028430
3  2012 male   0 to 4    2007742
```

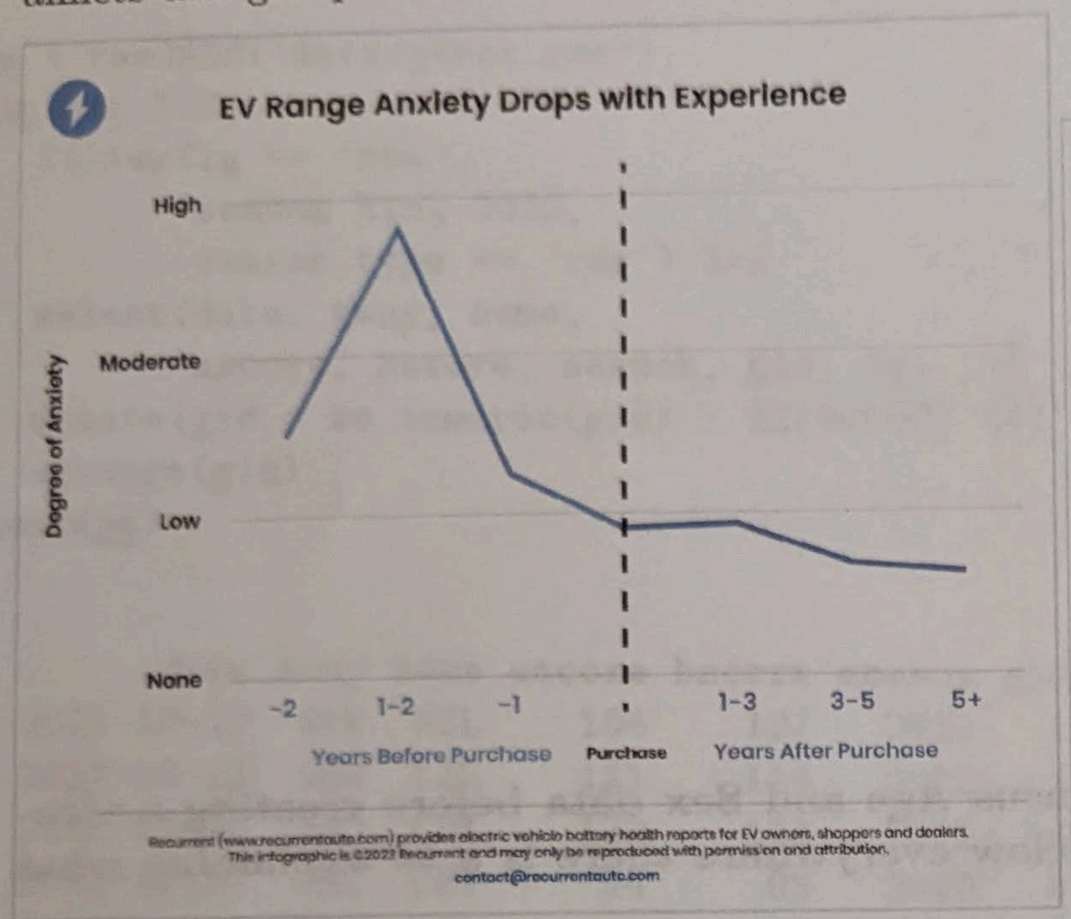
```
# A tibble: 3 x 4
   year mf     age.group   pop
  <dbl> <chr>  <fct>     <dbl>
1  2017 female 85+       67113
2  2018 female 85+       72153
3  2019 female 85+       76850
```

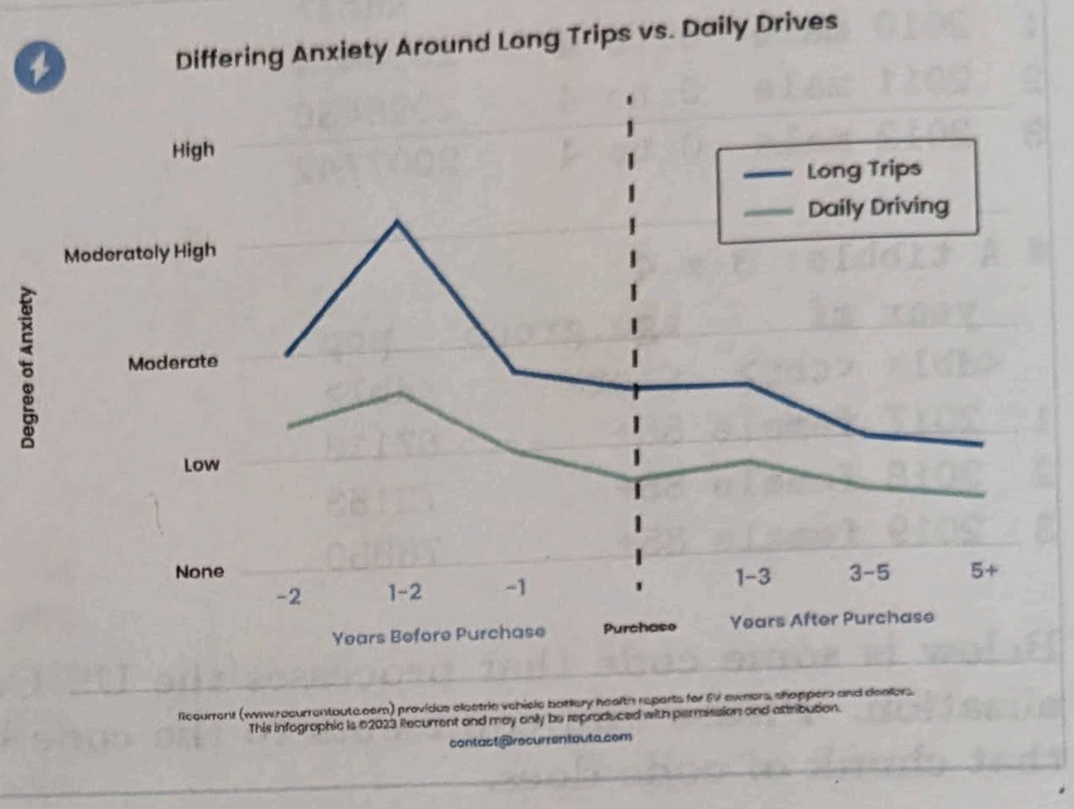
Below is some code that processes the US Census Age and Sex data before creating a visualization. Please add comments to the code below everywhere there is a ## explaining what that chunk of code does.

```
dg = d %>%

  ## remove years except 2010 and 2019 (first and last year)
  ##

  filter(year %in% c(2010, 2019)) %>%

  ##
  ## Convert year to factor so it's not numeric,

  mutate(year = factor(year)) %>%

  ## ~~Group~~ For each combo of year, age.group, and ~~mf~~ sex,
  ##
  ## find total population in that group
  group_by(year,
           age.group,
           mf) %>%
  summarise(pop = sum(pop))

  ## Plot population by age.group, with different windows
  ## for sex. Color by year 2010 or 2019

ggplot(dg,
       aes(x = pop,
           y = age.group,
           fill = year)) +
  geom_col(position = position_dodge()) +
  facet_wrap(~mf)
```

Visualization: EV Range Anxiety

The following figures appeared in the article https://www.recurrentauto.com/research/ev-range-anxiety-afflicts-this-group-most.





As summarized in the article, "Range anxiety refers to the fear that an electric car won't have enough juice [battery power] to reach its destination."

In this question you'll critique this visualization, discussing which best practices it follows and does not follow. Please give short answers (1 sentence, 25 words or less) to the questions below. Note that some of these questions are subjective and there aren't always necessarily "right" answers.

a. Do the title, axis labels, and other text clearly summarize the contents of the visualization? Why? Caption is super small. The x-axis labels are a bit odd (-2 1-2 -1). The y-axis are different for some reason (4 categories, then 5 categories. Other stuff seems fine

b. Does this visualization makes good use of color? Why?

Colors in first one are too similar (tough to tell difference on printed version)

c. Does the visualization give context, show what is "normal" or "average", or give some other point of reference? Why?

Would be nice to compare with ICE (internal combustion engine) vehicles

d. What would you change about the visualization? Include at least one additional comment different from your response given above.

dots on the lines, or change to bar chart

e. What is something interesting you learn about the data from this visualization? What do you think is one of the main takeaways?

Range anxiety drops a year before they purchase

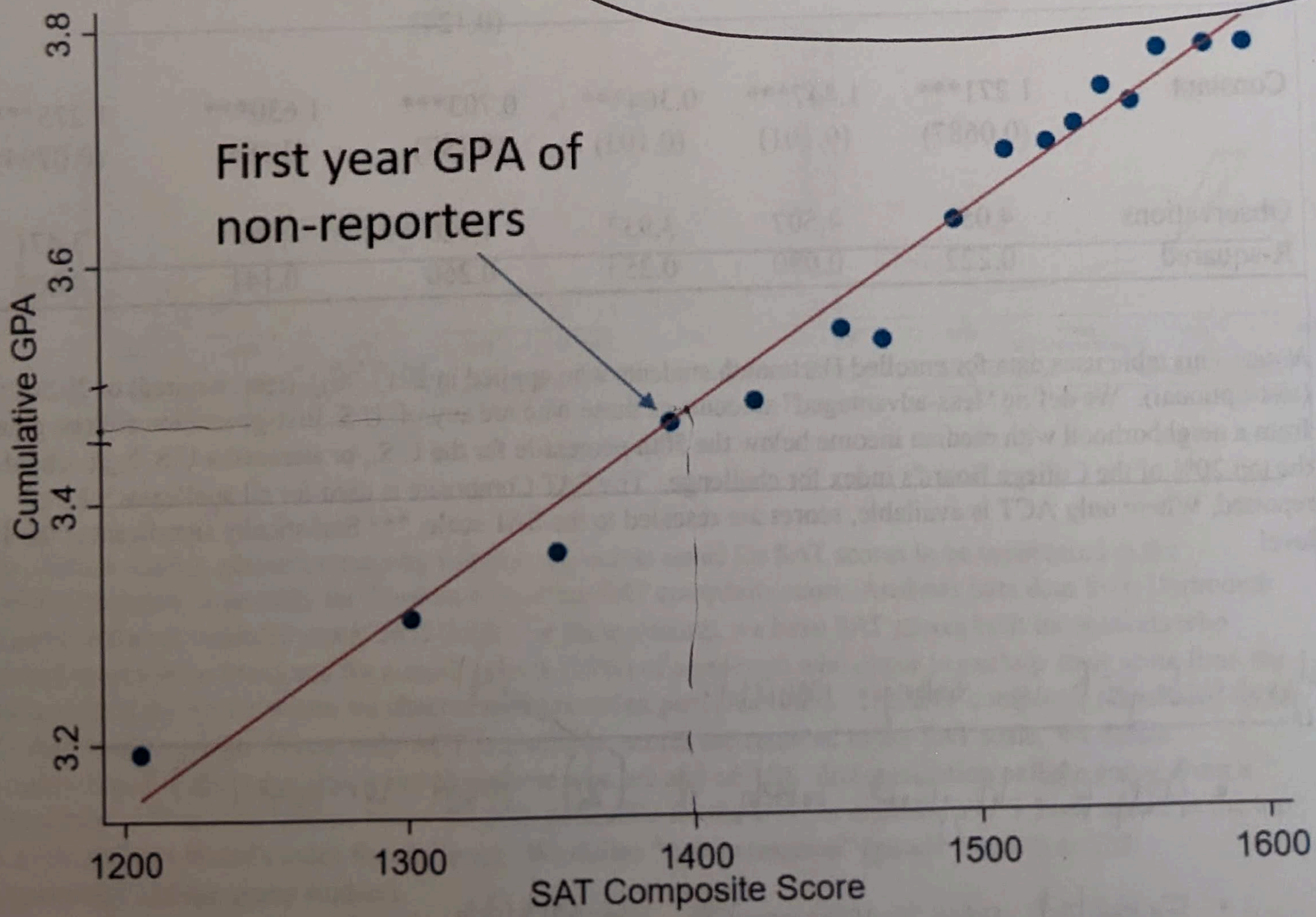
Interpreting visualizations and regression outputs

On January 30, 2024, Dartmouth released a report from a working group that was tasked to investigate the role of standardized test scores (SAT and ACT) in undergraduate admissions at Dartmouth. The difficulties of testing during COVID, coupled with known or perceived biases in standardized tests, led many schools to stop requiring that students submit their standardized test scores as part of the admissions process. Dartmouth was considering reinstating the requirement, but wanted empirical evidence that these test scores were actually helpful in the admissions process, and also hurting less-advantaged students. Among the data the working group had at their disposal were First Year GPA at Dartmouth, SAT scores, high school GPA, and high school rank. Give short (1 sentence, 25 words or less) answers to the questions below.

a. The following figure shows a bin scatter plot of First Year GPA vs SAT for 16 equal sized bins. The expected GPA for a student with an SAT score of 1400 looks to be about 3.5. The points look like on average that they are less than 0.05 GPA away from the line shown. Would you expect that most students with a 1400 SAT score would have a first year GPA in the range of 3.4-3.6? Why or why not?

No, [illegible] is approximately the CI (prediction interval would be much wider 3.4-3.6)

Figure 1.
Relationship Between Cumulative First-Year GPA and Composite SAT Scores:
Dartmouth Students



Notes: Figure displays bin scatter plot of cumulative first-year GPA against the SAT for 16 equal-sized bins of SAT for enrolled Dartmouth students in the 2017-2018 (test-required) and 2020-2021 (test-optional) cohorts. The SAT Composite is used for all applicants where reported. Where only ACT is available, scores are rescaled to the SAT scale. Non-reporters have average GPAs at the 31st percentile of cumulative GPA.

b. As part of their analysis, the group fit several regression models. A table of results is given below. Based on these results, do you recommend that Dartmouth use SAT scores as part of their admissions process?

Table 1.
Value of High School GPA and the SAT Composite in Predicting Cumulative First-Year GPA at Dartmouth

	(1) First Year GPA	(2) First Year GPA	(3) First Year GPA	(4) First Year GPA	(5) First Year GPA Less Adv	(6) First Year GPA More Adv
SAT Composite Score	0.00158***		0.00139***	0.00140***	0.00129***	0.00158***
	(4.64e-05)		(4.88e-05)	(6.56e-05)	(0.000132)	(5.31e-05)
HS GPA		0.550***	0.322***	0.220***		
		(0.0261)	(0.0263)	(0.0485)		
Class Rank				-0.490***		
				(0.124)		
Constant	1.271***	1.447***	0.304***	0.703***	1.630***	1.275***
	(0.0687)	(0.101)	(0.103)	(0.197)	(0.185)	(0.0794)
Observations	4,051	4,507	3,937	1,920	580	3,471
R-squared	0.222	0.090	0.255	0.260	0.141	0.203

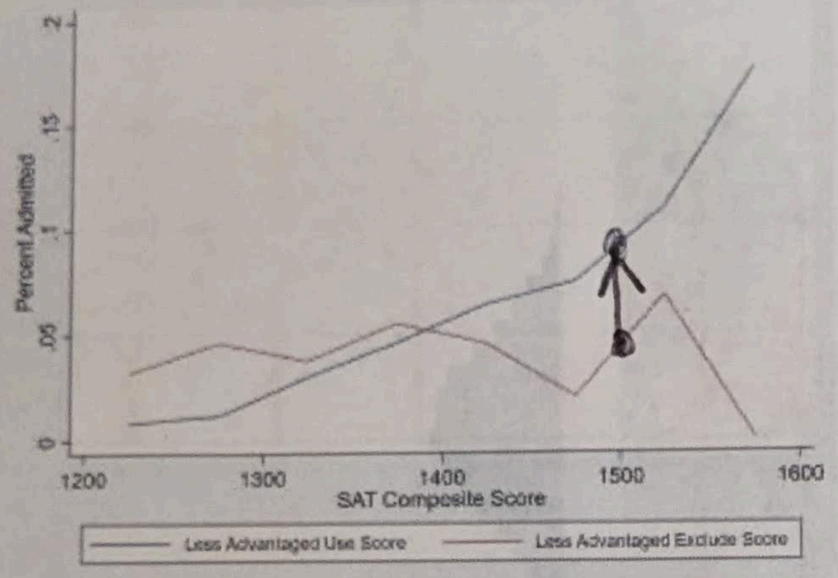
Notes: This table uses data for enrolled Dartmouth students who applied in 2017-2018 (test-required) or 2021-2022 (test-optional). We define "less-advantaged" students as those who are any of: U.S. first-generation college going, from a neighborhood with median income below the 50th percentile for the U.S., or attended a U.S. high school in the top 20% of the College Board's index for challenge. The SAT Composite is used for all applicants where reported. Where only ACT is available, scores are rescaled to the SAT scale. *** Statistically significant at the 1% level

Yes.
- Signif in every model.
- Adj R^2 higher with it (3) or (4) than without (2)
- Expected sign, reasonable magnitude

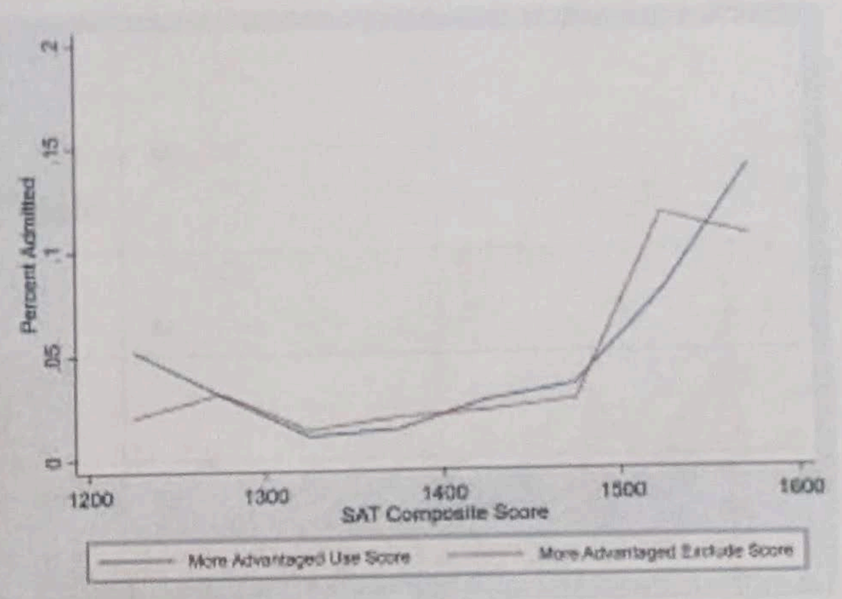
c. According to Figure 6 below, if a less-advantaged student has an SAT Composite Score of 1500, should they submit their score as part of the admissions process when given the option?

Figure 6.
Admissions Rates of Dartmouth Applicants Submitting and Not Submitting Scores in the Test Optional Cohorts, by SAT Score: By Advantage and First-Generation Status

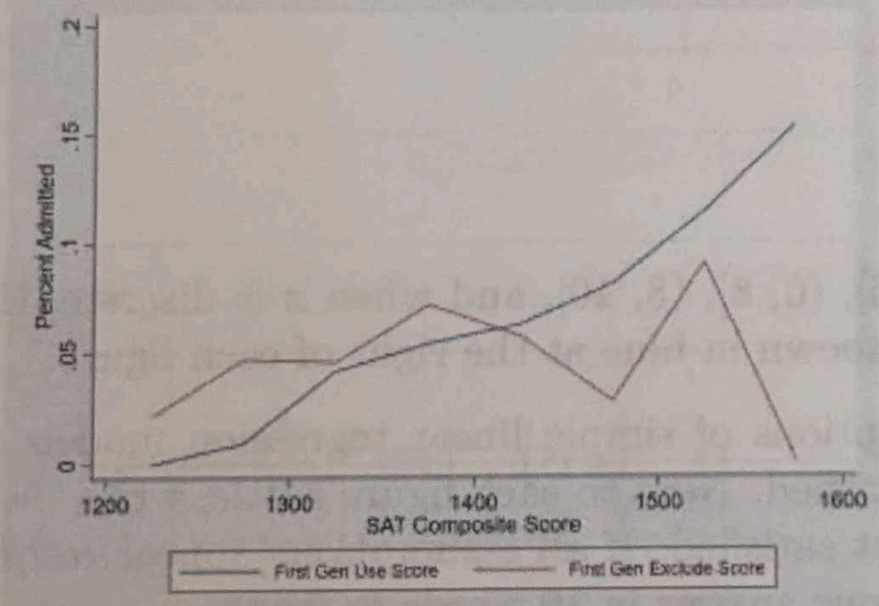
a. Less-Advantaged Applicants



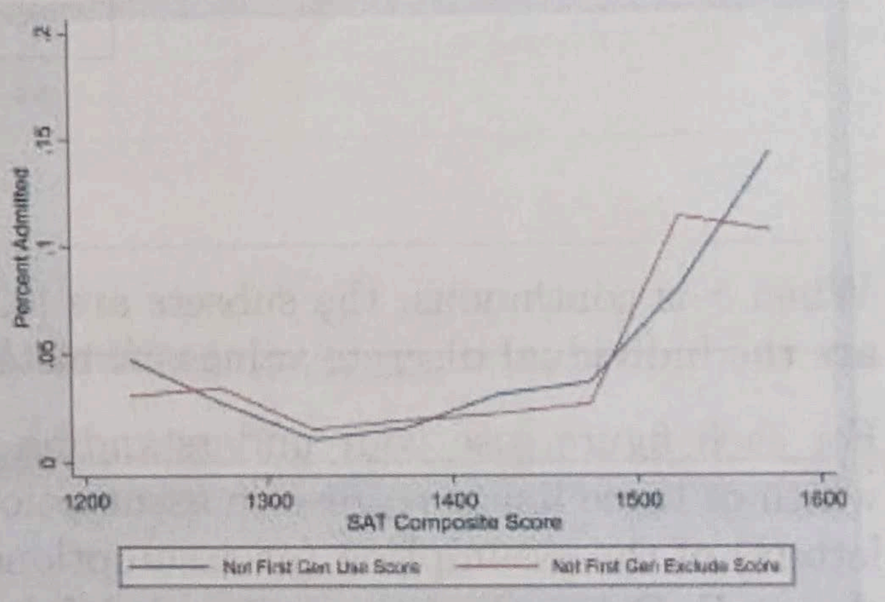
b. More-Advantaged Applicants



c. First-Generation Applicants



d. Not First-Generation Applicants



Notes: Charts display admission rates by whether applicants opted for SAT scores to be considered in the application decision, separately for 50-point bins of the SAT composite score. Analysis uses data from Dartmouth applicants in the test-optional years, 2021-2022. For these cohorts, we have SAT scores both for students who submitted scores (blue lines) and for a small sample (19%) of applicants who chose to exclude their score from the admission decision but for whom we observe their scores ex post (red lines). The SAT composite is included for all applicants where reported. Where only ACT is available, scores are rescaled to the SAT scale. We define "less-advantaged" students (panels a and b) as those who are any of: U.S. first-generation college going, from a neighborhood with median income below the 50th percentile for the U.S., or attended a U.S. high school in the top 20% of the College Board's index for challenge. We define "first generation" (panels c and d) as U.S. first-generation college-going students.

Yes, percent admitted goes from 0.05 to 0.10.

Figure 1. Answer: E

Explanation: - y appear roughly normal for subsets
- constant variance
- dots are linear

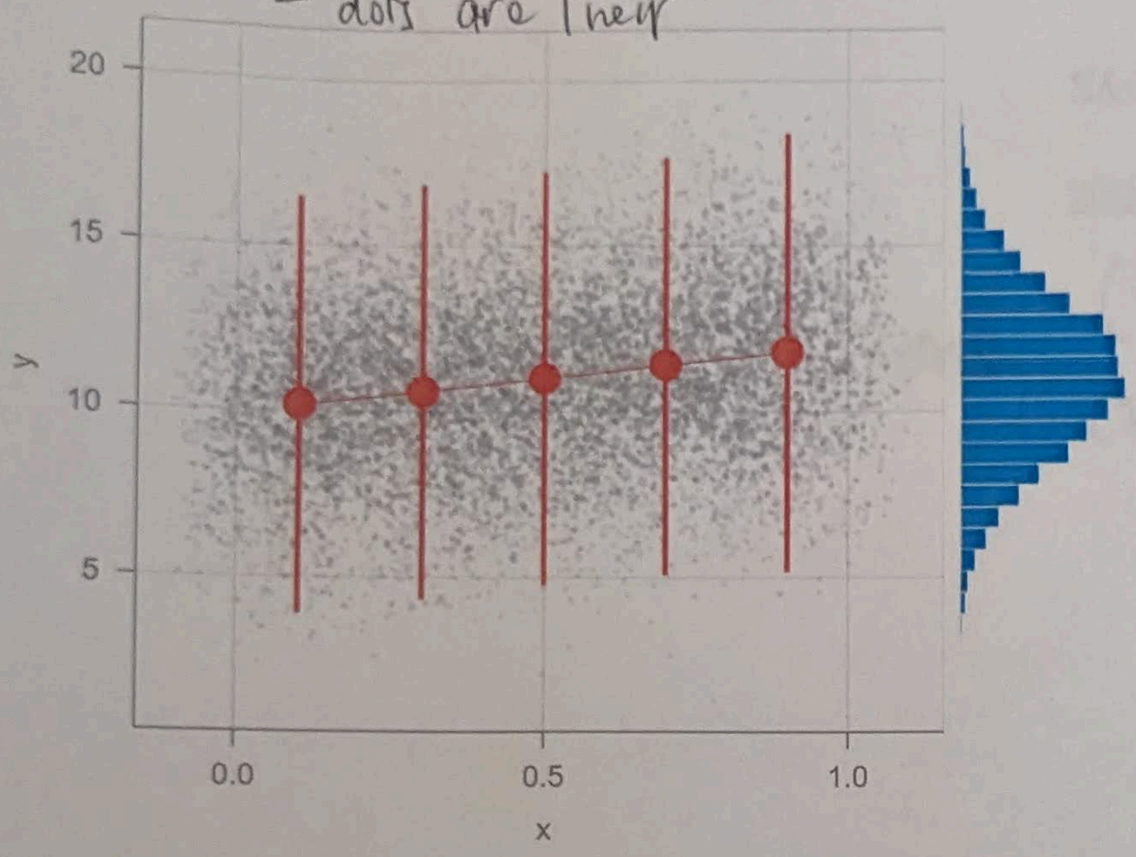


Figure 2. Answer: A, B, C

Explanation: y right skewed
exponential relationship
variance increasing

) do a log transform

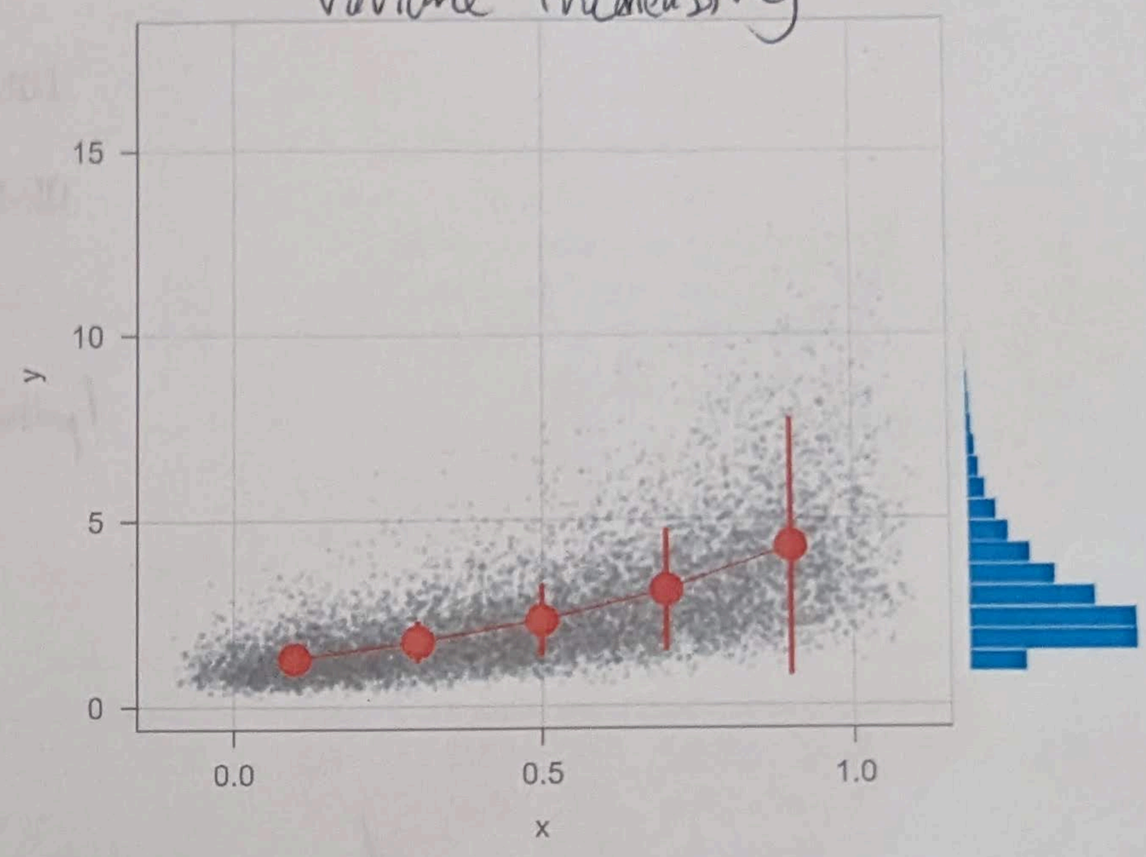


Figure 3. Answer: E

Explanation: y is not normal, but that is irrelevant. y is normal for each x in this case.

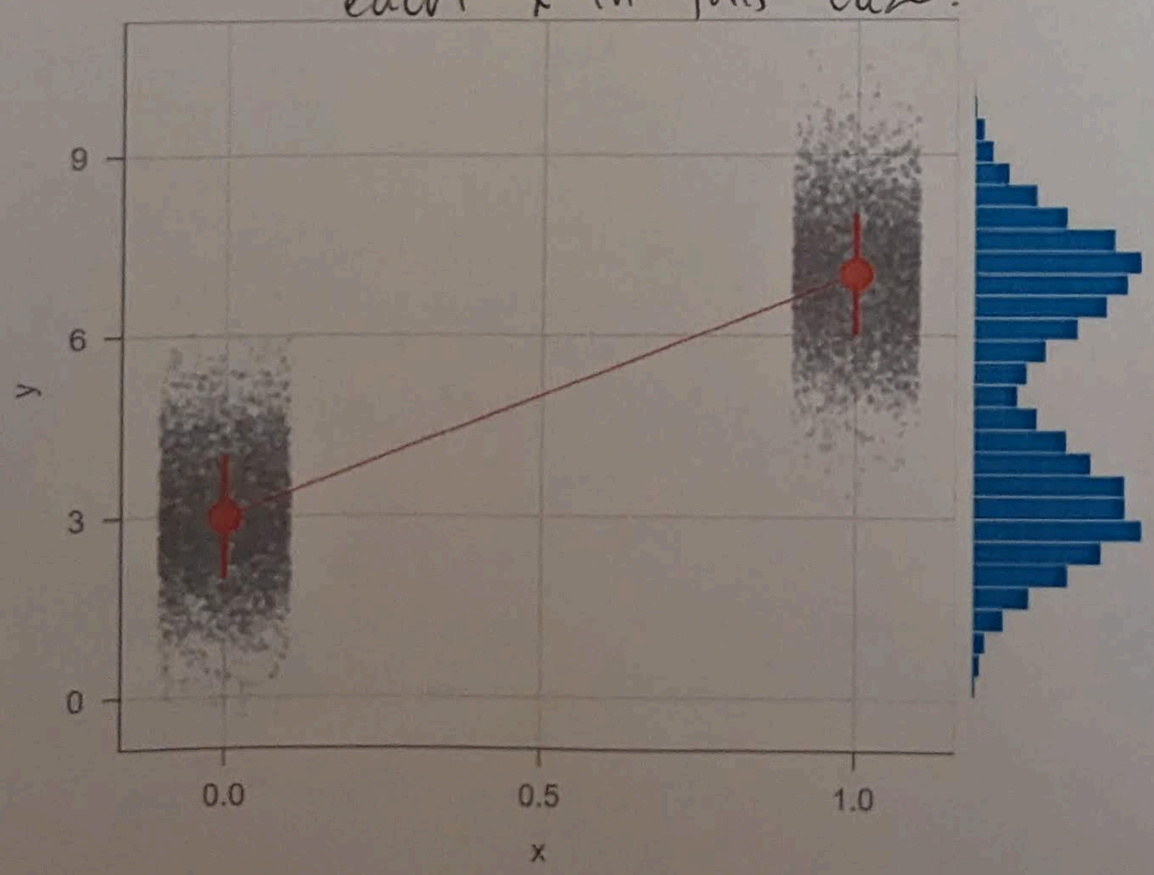


Figure 4. Answer: B

Explanation: y is discrete, not normal