Midterm on 6.20

- Scope (Lec1—Lec14): Probability, Sampling, Pandas, Regex, Visualization, KDE, Loss Function, Simple Linear Regression, Feature Engineering, Cross Validation and Regularization

  (possibly taught in week 6)

- DO NOT forget to go over the basic concepts

e.g. sampling methods, different types of bias, or even the definition of data science!

Midterm:

- 1-hr long
- composed of multiple choice & coding questions
- open-book, open-note

IMPORTANT:

- There may be a great number of questions, so please arrange your time wisely. If you spend too much time reviewing slides or checking your answer using Python code, you may not be able to finish all the questions.
- You are encouraged to work the coding questions out. Partial credits will be given even if your answer is not completely right.

Key parts: GroupBy, Pivot Table, String Operations…

Sampling

[2 Pts] A bootstrap sample consists of n draws made uniformly at random with replacement from an original set of n individuals.

Person A is an individual in the original set. The chance that Person A appears in the bootstrap sample is:

○ $\frac{1}{n}$ ○ $1 - \frac{1}{n}$ ○ $\left(\frac{1}{n}\right)^n$ ○ $1 - \left(\frac{1}{n}\right)^n$ ○ $1 - \left(1 - \frac{1}{n}\right)^n$

~~(D)~~ E

A university wants to study the experience of students enrolled in its big classes, defined as classes with enrollments of 500 or more. There are 20 such classes. From each of these classes, one enrolled student is chosen uniformly at random to take part in the university's survey. You can assume that the selection from each class is performed independently of the selections in the other classes. In this scenario:

(a) [1 Pt] There are students in the population of interest who are not in the sampling frame.

○ True ○ False

(b) [1 Pt] There are students in the sampling frame who are not in the population of interest.

○ True ○ False

(c) [1 Pt] The method of sampling produces a probability sample of students enrolled in the big classes.

○ True ○ False

(d) [1 Pt] The method of sampling produces a simple random sample of students enrolled in the big classes.

○ True ○ False

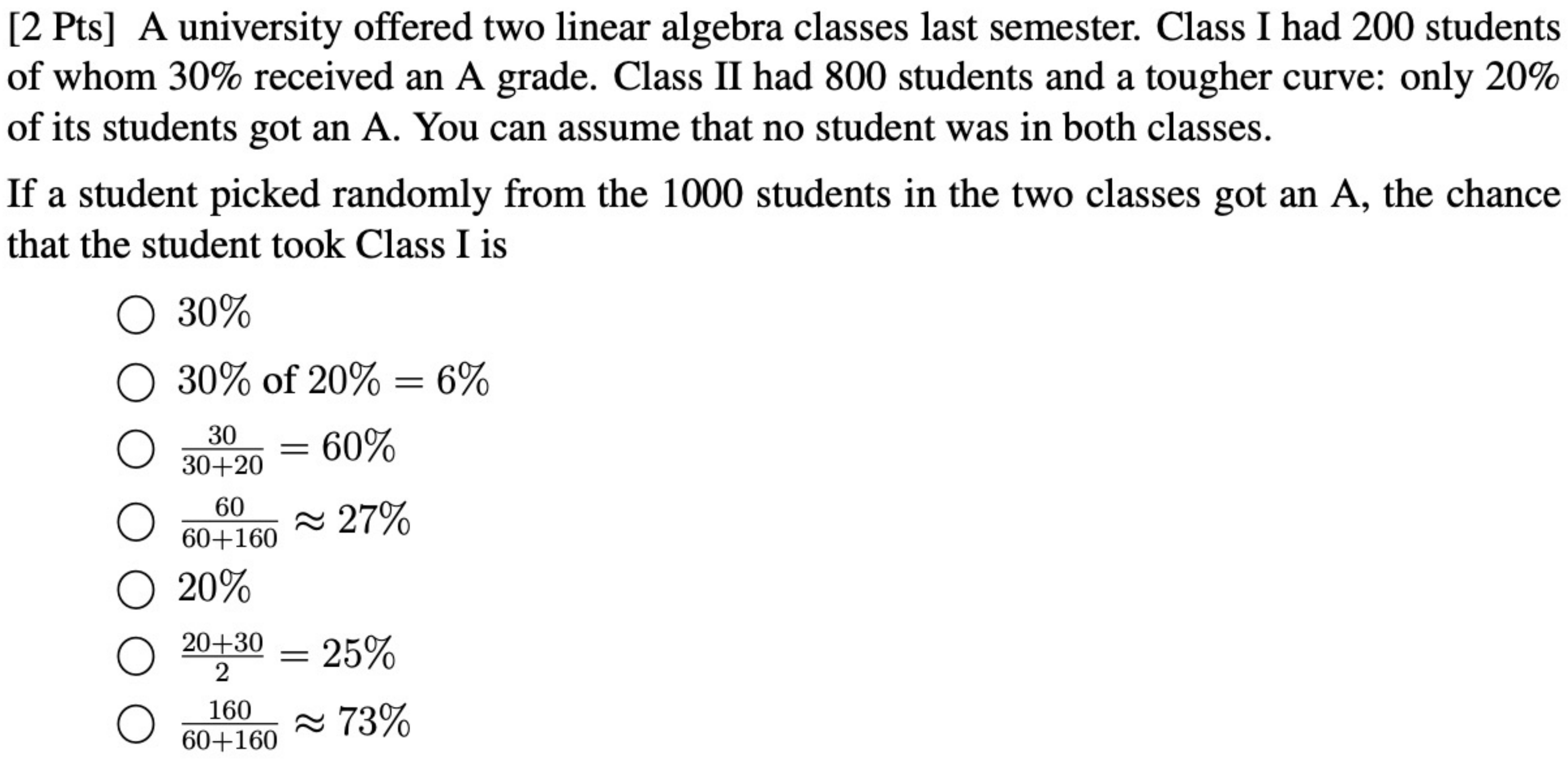
(e) [1 Pt] Because a student is chosen from each class, all students in the big classes have the same chance of being selected.

○ True ○ False

F/F/T/F/F/F

(f) [1 Pt] Because a student is chosen from each of 20 big classes, there will be 20 students in the sample.

○ True ○ False



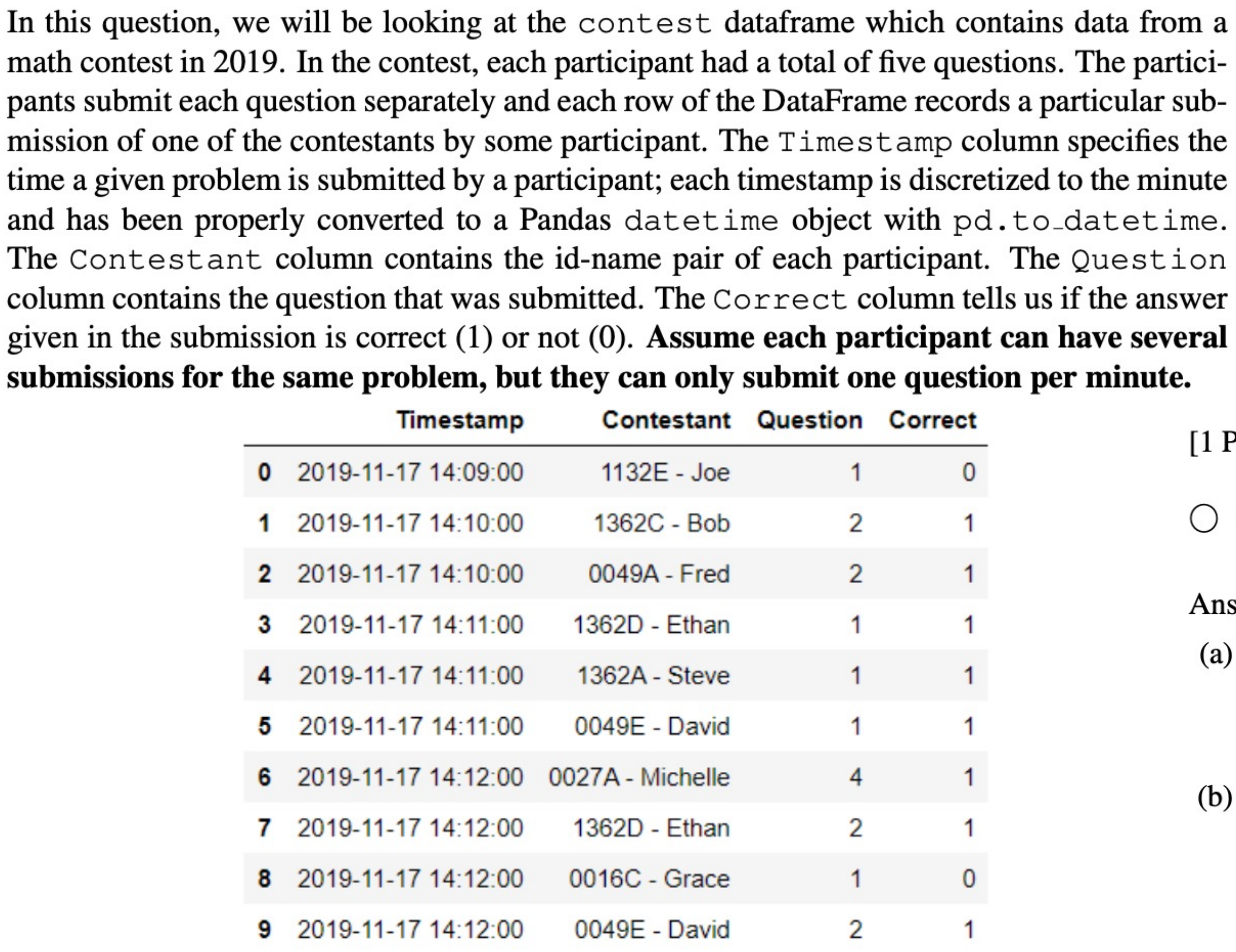
[2 Pts] A university offered two linear algebra classes last semester. Class I had 200 students of whom 30% received an A grade. Class II had 800 students and a tougher curve: only 20% of its students got an A. You can assume that no student was in both classes.

If a student picked randomly from the 1000 students in the two classes got an A, the chance that the student took Class I is

- ○ 30%
- ○ 30% of $20\% = 6\%$
- ○ $\frac{30}{30+20} = 60\%$
- ○ $\frac{60}{60+160} \approx 27\%$
- ○ 20%
- ○ $\frac{20+30}{2} = 25\%$
- ○ $\frac{160}{60+160} \approx 73\%$

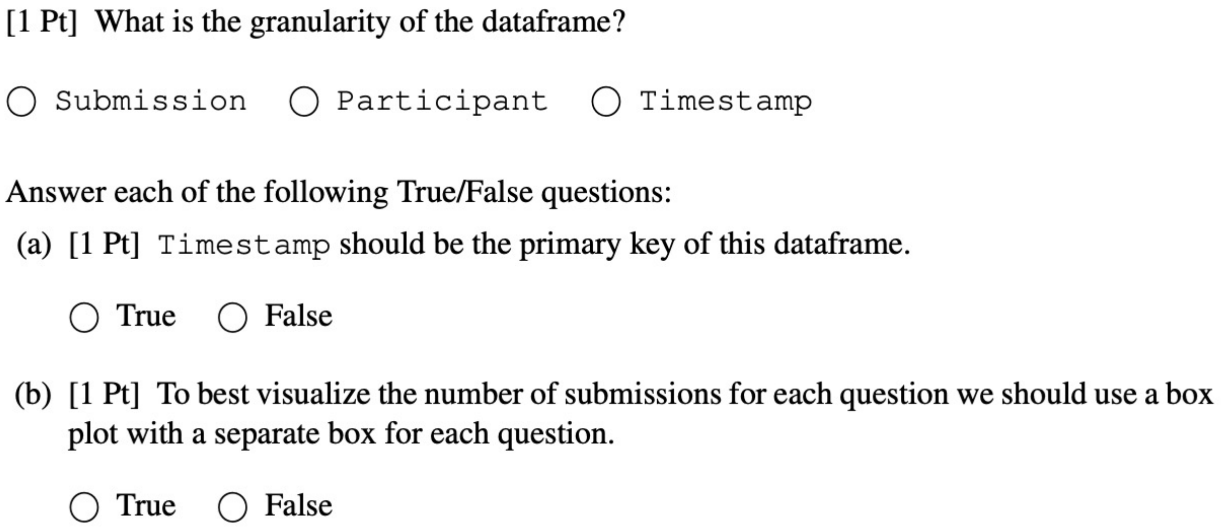
(D)

Dataframe and pandas



In this question, we will be looking at the `contest` dataframe which contains data from a math contest in 2019. In the contest, each participant had a total of five questions. The participants submit each question separately and each row of the DataFrame records a particular submission of one of the contestants by some participant. The `Timestamp` column specifies the time a given problem is submitted by a participant; each timestamp is discretized to the minute and has been properly converted to a Pandas `datetime` object with `pd.to_datetime`. The `Contestant` column contains the id-name pair of each participant. The `Question` column contains the question that was submitted. The `Correct` column tells us if the answer given in the submission is correct (1) or not (0). **Assume each participant can have several submissions for the same problem, but they can only submit one question per minute.**

	Timestamp	Contestant	Question	Correct
0	2019-11-17 14:09:00	1132E - Joe	1	0
1	2019-11-17 14:10:00	1362C - Bob	2	1
2	2019-11-17 14:10:00	0049A - Fred	2	1
3	2019-11-17 14:11:00	1362D - Ethan	1	1
4	2019-11-17 14:11:00	1362A - Steve	1	1
5	2019-11-17 14:11:00	0049E - David	1	1
6	2019-11-17 14:12:00	0027A - Michelle	4	1
7	2019-11-17 14:12:00	1362D - Ethan	2	1
8	2019-11-17 14:12:00	0016C - Grace	1	0
9	2019-11-17 14:12:00	0049E - David	2	1



[1 Pt] What is the granularity of the dataframe?

○ `Submission` ○ `Participant` ○ `Timestamp`

Answer each of the following True/False questions:

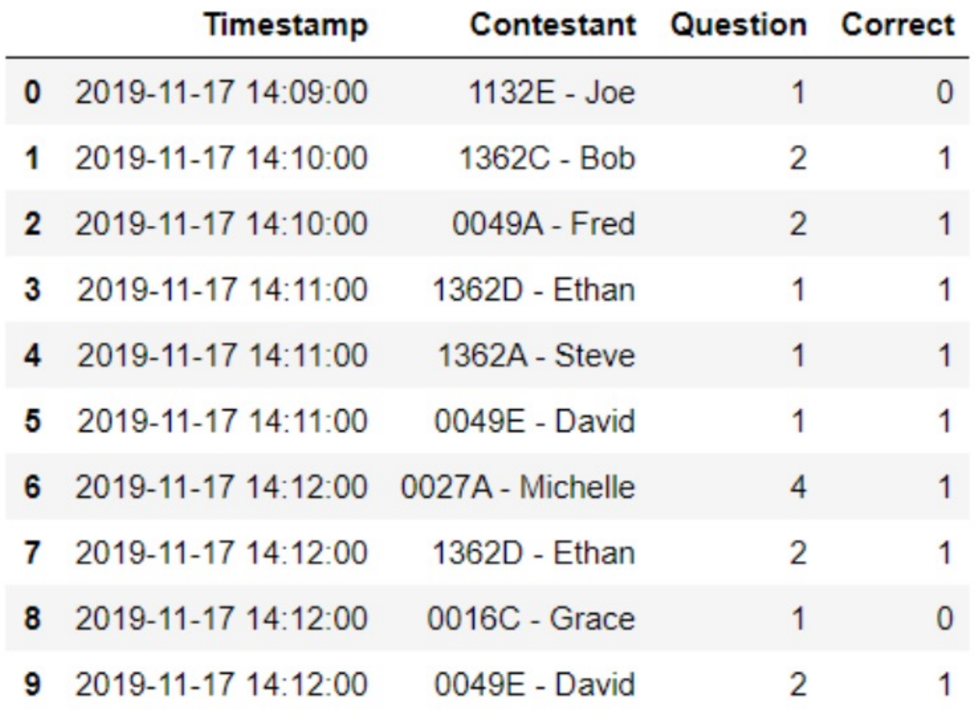
(a) [1 Pt] `Timestamp` should be the primary key of this dataframe.

○ True ○ False

(b) [1 Pt] To best visualize the number of submissions for each question we should use a box plot with a separate box for each question.

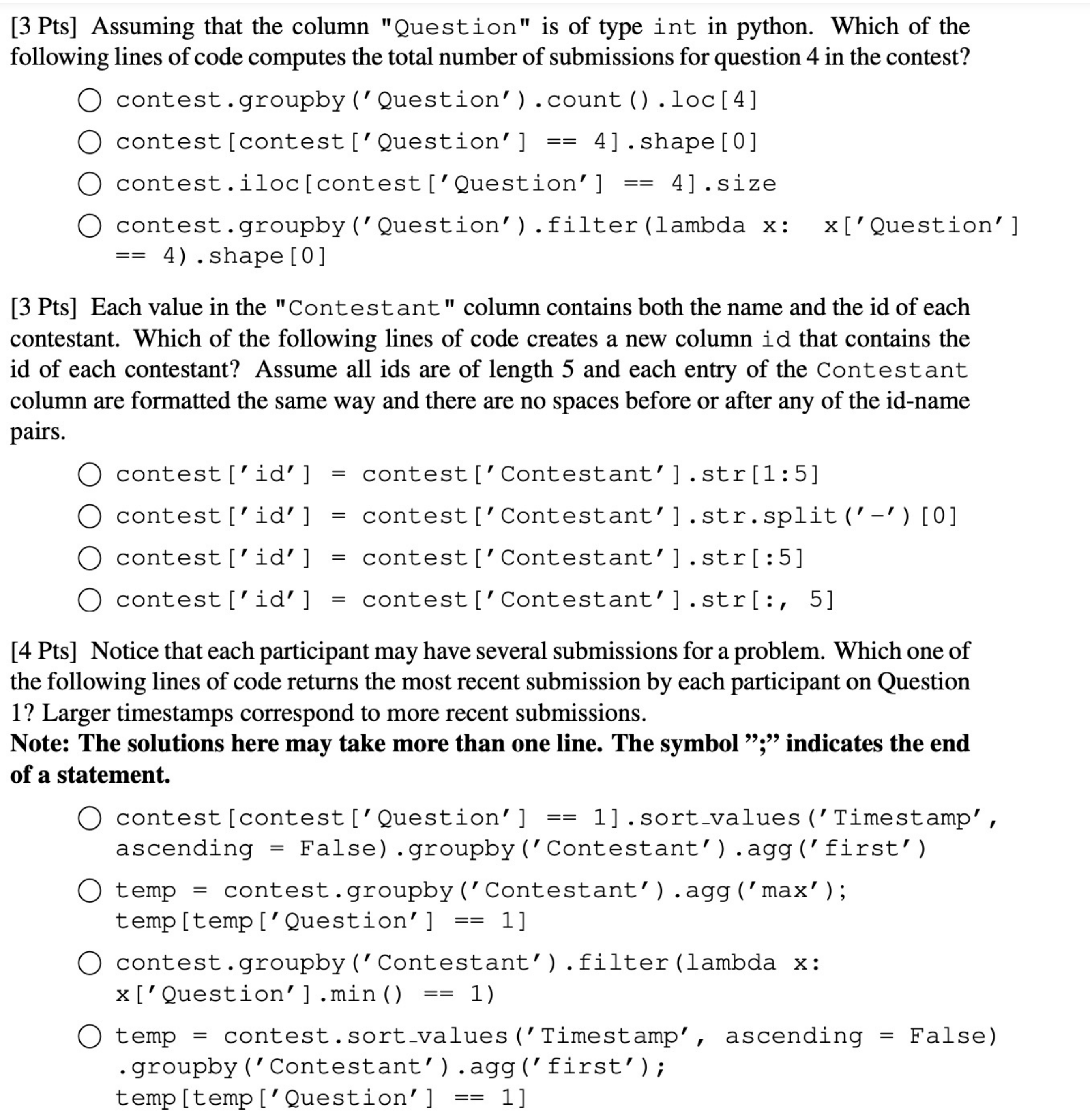
○ True ○ False

A/F/F



	Timestamp	Contestant	Question	Correct
0	2019-11-17 14:09:00	1132E - Joe	1	0
1	2019-11-17 14:10:00	1362C - Bob	2	1
2	2019-11-17 14:10:00	0049A - Fred	2	1
3	2019-11-17 14:11:00	1362D - Ethan	1	1
4	2019-11-17 14:11:00	1362A - Steve	1	1
5	2019-11-17 14:11:00	0049E - David	1	1
6	2019-11-17 14:12:00	0027A - Michelle	4	1
7	2019-11-17 14:12:00	1362D - Ethan	2	1
8	2019-11-17 14:12:00	0016C - Grace	1	0
9	2019-11-17 14:12:00	0049E - David	2	1

B/C/A



[3 Pts] Assuming that the column "Question" is of type int in python. Which of the following lines of code computes the total number of submissions for question 4 in the contest?

○ contest.groupby('Question').count().loc[4]

○ contest[contest['Question'] == 4].shape[0]

○ contest.iloc[contest['Question'] == 4].size

○ contest.groupby('Question').filter(lambda x: x['Question'] == 4).shape[0]

[3 Pts] Each value in the "Contestant" column contains both the name and the id of each contestant. Which of the following lines of code creates a new column id that contains the id of each contestant? Assume all ids are of length 5 and each entry of the Contestant column are formatted the same way and there are no spaces before or after any of the id-name pairs.

○ contest['id'] = contest['Contestant'].str[1:5]

○ contest['id'] = contest['Contestant'].str.split('-')[0]

○ contest['id'] = contest['Contestant'].str[:5]

○ contest['id'] = contest['Contestant'].str[:, 5]

[4 Pts] Notice that each participant may have several submissions for a problem. Which one of the following lines of code returns the most recent submission by each participant on Question 1? Larger timestamps correspond to more recent submissions.
Note: The solutions here may take more than one line. The symbol ";" indicates the end of a statement.

○ contest[contest['Question'] == 1].sort_values('Timestamp', ascending = False).groupby('Contestant').agg('first')

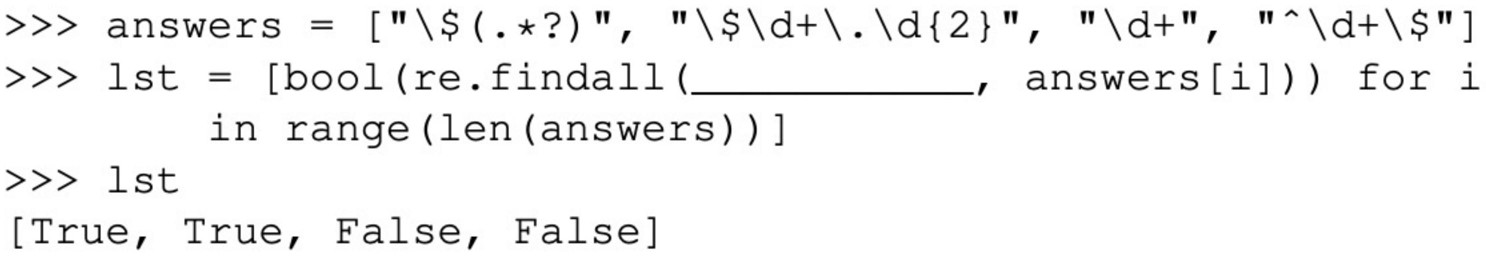
○ temp = contest.groupby('Contestant').agg('max');
temp[temp['Question'] == 1]

○ contest.groupby('Contestant').filter(lambda x: x['Question'].min() == 1)

○ temp = contest.sort_values('Timestamp', ascending = False)
.groupby('Contestant').agg('first');
temp[temp['Question'] == 1]

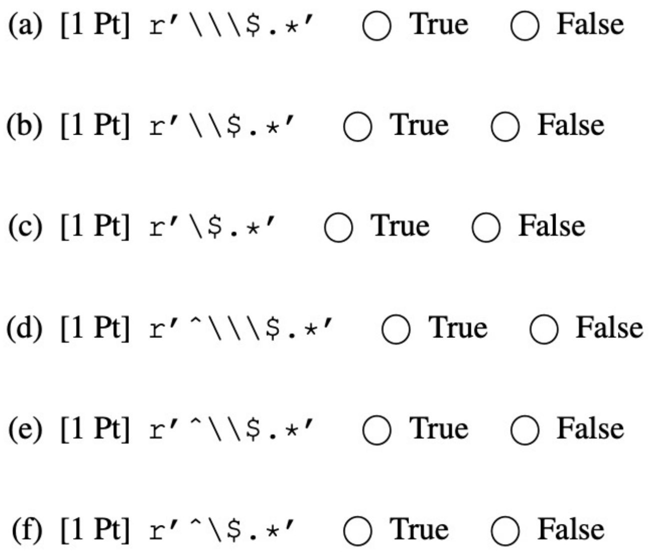
Regex

If you have designed a Regex to check for a dollar sign at the beginning of some regular expression. Which of the following regular expressions would work such that the following Python expression outputs correctly?



```
>>> answers = ["\$(.*?)", "\$\d+\.\d{2}", "\d+", "^\d+\$"]
>>> lst = [bool(re.findall(____________, answers[i])) for i
        in range(len(answers))]
>>> lst
[True, True, False, False]
```

?/F/F/T/F/F



(a) [1 Pt] r’\\\$.*’ ○ True ○ False

(b) [1 Pt] r’\\$.*’ ○ True ○ False

(c) [1 Pt] r’\$.*’ ○ True ○ False

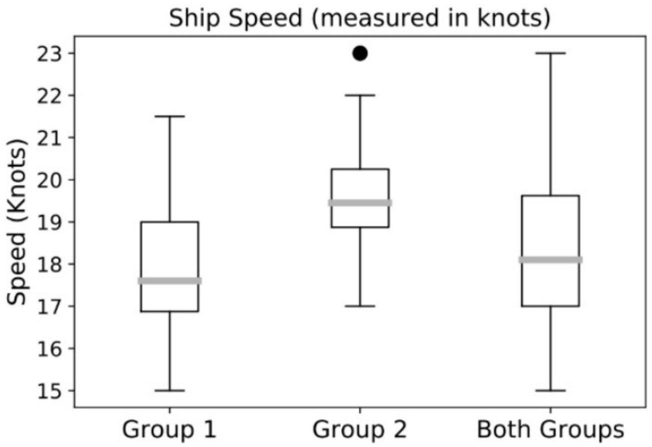
(d) [1 Pt] r’^\\\$.*’ ○ True ○ False

(e) [1 Pt] r’^\\$.*’ ○ True ○ False

(f) [1 Pt] r’^\$.*’ ○ True ○ False

Visualization

The plot below summarizes the distributions of speeds of two groups of ships. The box plot on the extreme right is for all the ships. The other two box plots are for the individual groups.



F/T/F/B

For parts a through c, consider the box plot for Group 1. Based on this box plot alone, what can you conclude about the % of speeds in Group 1 that are 18 knots or more?

(a) [1 Pt] $< 50\%$ of the speeds in Group 1 are 18 knots or more

○ True

○ False

(b) [1 Pt] $\leq 50\%$ of the speeds in Group 1 are 18 knots or more

○ True

○ False

(c) [1 Pt] $\geq 50\%$ of the speeds in Group 1 are 18 knots or more

○ True

○ False

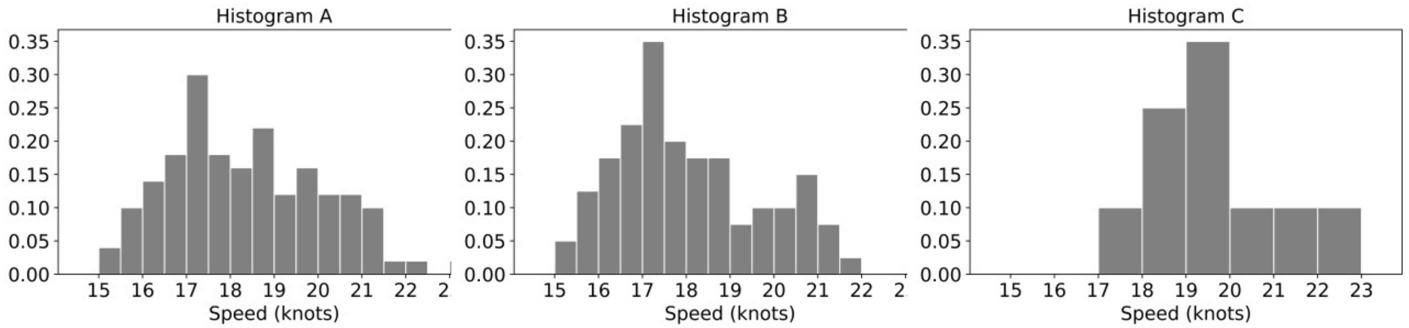
(d) [1 Pt] Which one of the following statements can we conclude from the boxplots?

○ Since the box in the Group 1 plot is bigger than the box in the Group 2 plot, we can conclude that there are more ships in Group 1 than in Group 2.

○ Since the distribution in the Both Groups plot is much closer to that of Group 1 than of Group 2, we can conclude that there are more ships in Group 1 than in Group 2.

○ Based on these box plots, it is not possible to determine whether there are more ships in Group 1 than in Group 2.

All of the histograms below are based on the box plots above, and are drawn to the density scale.



For parts a through c, match the histograms to their corresponding box plot.

(a) [1 Pt] Histogram A: ○ Group 1 ○ Group 2 ○ Both Groups

(b) [1 Pt] Histogram B: ○ Group 1 ○ Group 2 ○ Both Groups

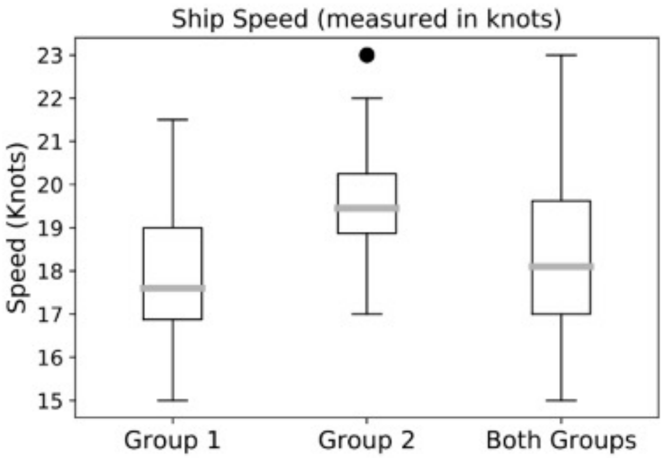
(c) [1 Pt] Histogram C: ○ Group 1 ○ Group 2 ○ Both Groups

(d) [1 Pt] About ______________% of the speeds in Histogram C are in the [17, 18) bin.

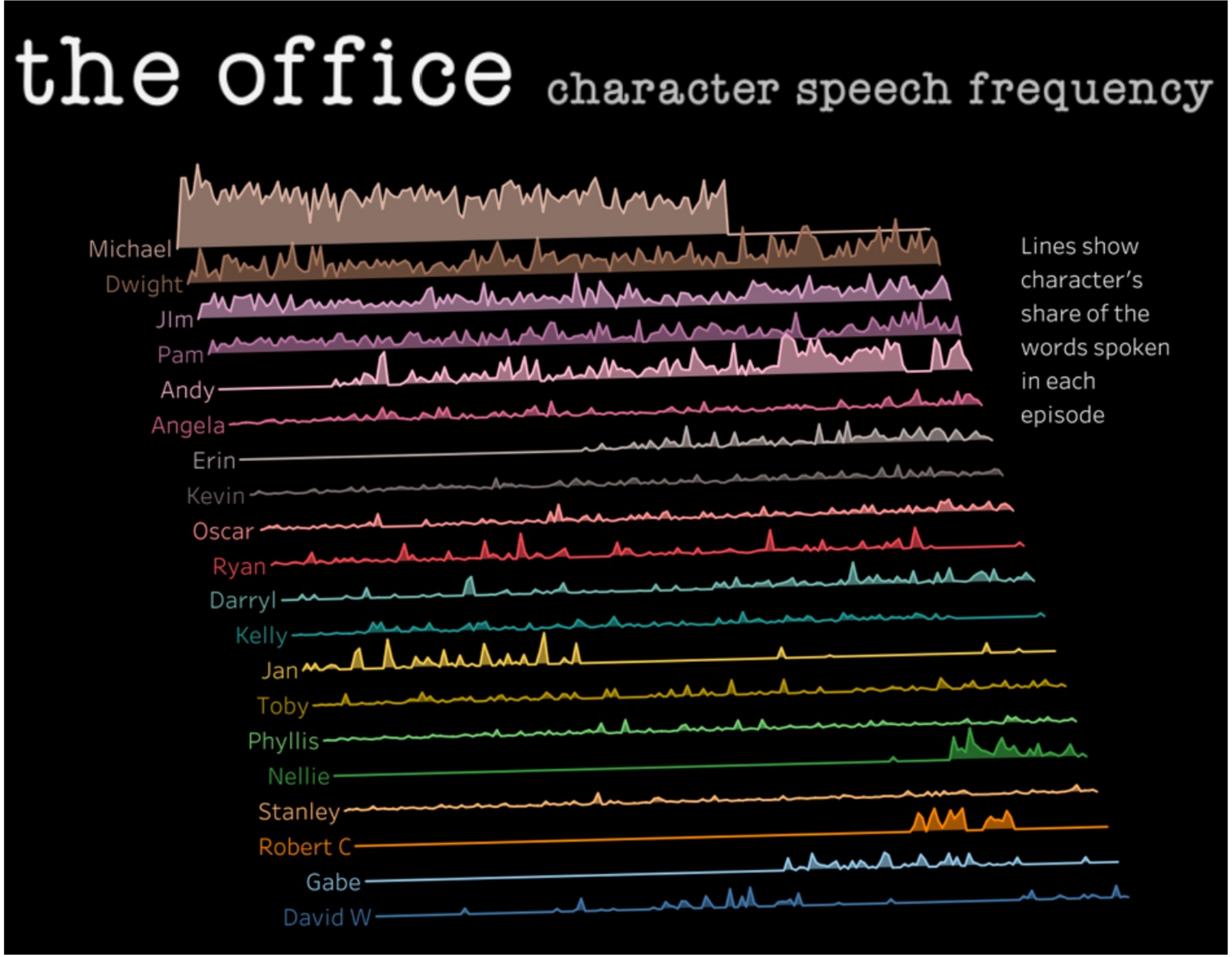
○ 5 ○ 10 ○ 15 ○ 20 ○ 25 ○ 30 ○ 35

(e) [1 Pt] In Histogram A, the percent of speeds in the [17, 18) range is:

○ between 30 and 50
○ between 20 and 25
○ less than 10



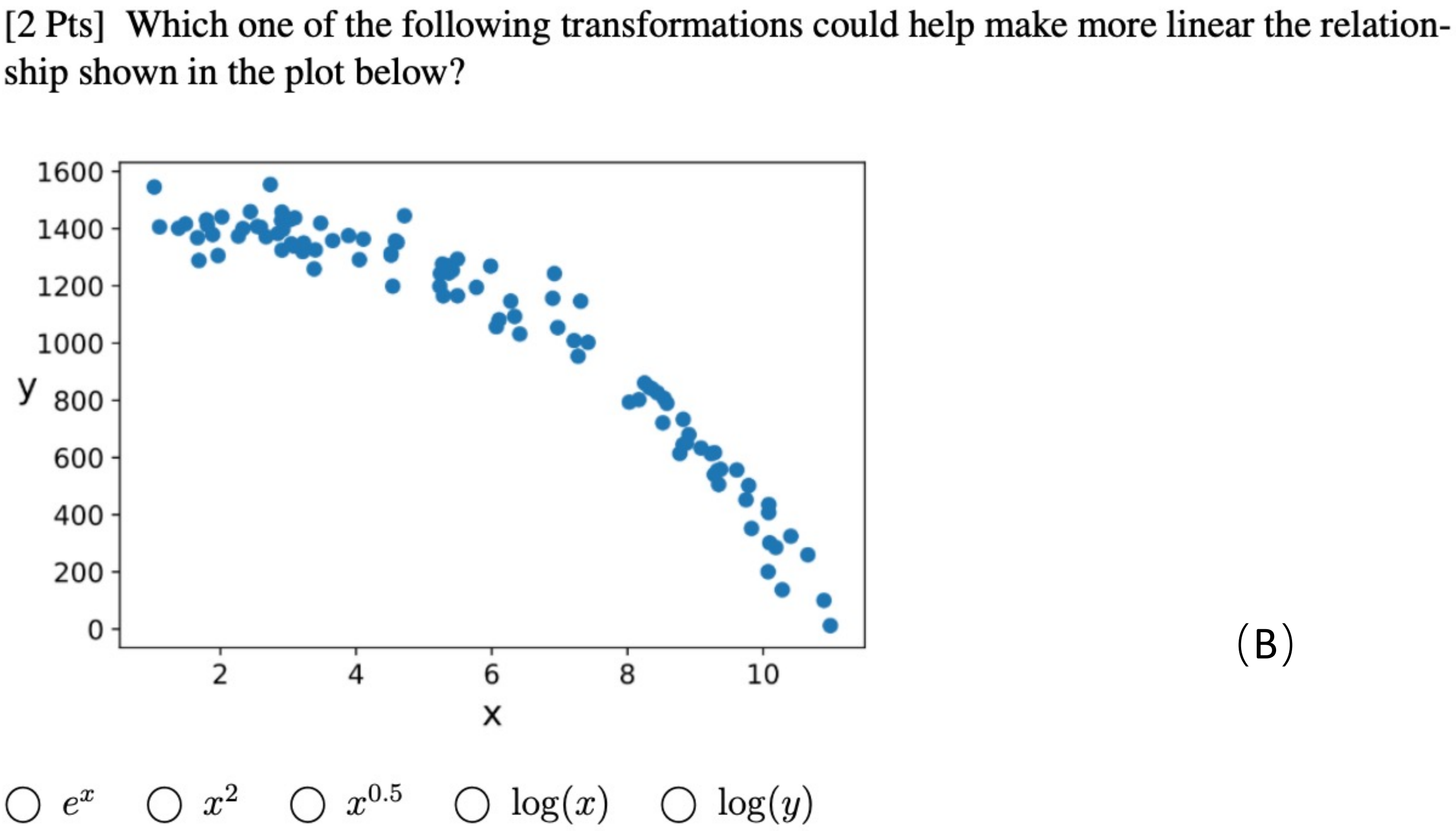
C/A/B/B/B



X-axis doesn't have a label
X-axis is not aligned

Name a flaw with the visualization:

Transformation



[2 Pts] Which one of the following transformations could help make more linear the relationship shown in the plot below?

○ e^x ○ x^2 ○ $x^{0.5}$ ○ $\log(x)$ ○ $\log(y)$

(B)