Data 8 Foundations of Data Science
Fall 2021

Final Exam

INSTRUCTIONS

- You must **write your name in the space provided on one side of every page of the paper exam.**
- The exam is worth 120 points. You have 170 minutes to complete it.
- An official final exam reference is provided. You may not use any other paper, reference, source, or computational device or system apart from those permitted for the online exam.
- Write/mark your answers on the exam in the blanks/bubbles/boxes provided. Answers written anywhere else will not be graded. Unless the question specifically asks you to explain your answer, you do not need to do so, and if you write an explanation it will not be graded.
- For all Python code, you may assume that the statements `from datascience import *` and `import numpy as np` have been executed. Do not use features of the Python language that have not been described in this course.
- In any part, you are free to use any tables, arrays, or functions that have been defined in previous parts of the same question, and you may assume they have been defined correctly.
- In starter code we provide, ____________ can mean any code, including commas and periods.

Last name	
First name	
Student ID number	
Calcentral email (`_@berkeley.edu`)	
Lab GSI	
Your room number and building (e.g. 1 Pimentel)	
Your seat number (e.g. A1)	
← Name of the person to your left	
Name of the person to your right →	
All the work on this exam is my own. **(please sign)**	

1. (3 points) Counting

Define a function `count_elem` that takes two arguments: an array `a` and a value `x`. It should return the number of times that `x` appears as an element of `a`.

For example, `count_elem(make_array('cat', 'cat', 'dog'), 'cat')` should return 2.

```
def count_elem(a, x):

    ____________________ np.________________________________(____________________)
```

2. (15 points) National Parks

After helping students during office hours with the Old Faithful lab, Raymond was interested in learning more about the famous geyser. After some quick research, he discovered that Old Faithful was in Yellowstone National Park. His curiosity was piqued, so he decided to do more research on other national parks. He found the following table `parks` and wanted to answer some questions, but he needs your help! For each of the following questions, write a line of code that will answer his question. The `parks` table is shown below.

Name	Location	Date established	Area	Entrance fee	Visitors
Acadia	Maine	1919	49000	25	3000000
Yosemite	California	1890	761000	35	2268000
Arches	Utah	1971	76000	30	1600000

...(28 more rows)

(a) (3 pt) In which year was the first national park appearing in the `parks` table established?

____

(b) (3 pt) Assume each visitor to a park pays the corresponding entrance fee. Create a new table called `with_revenue` that contains all columns from the original `parks` table, plus a new column called `'Revenue'` that shows how much money each park collected in entrance fees for that year (number of visitors times the entrance fee).

`with_revenue =` ____

(c) (3 pt) Some of the national parks in the US are also designated as UNESCO World Heritage Sites, which are sites of importance to cultural or natural heritage. The table `unesco`, shown below, provides a list of national parks that are also UNESCO World Heritage Sites. How many national parks located in California are also designated as UNESCO World Heritage Sites?

Park
Yosemite
Glacier
Olympic
Everglades

...(54 more rows)

____

(d) **(3 pt)** After looking at the `parks` table again, Raymond realized that it may be easier to interpret the geographical size of each park by assigning it one of the labels `"Small"`, `"Medium"`, or `"Large"`.

Using the skeleton code below, write a function that takes in a numeric area as input and returns a string corresponding to the geographical sizing group it belongs to. Use the following table for reference:

Category	**Area Range**
Small	[0, 100000)
Medium	[100000, 500000)
Large	[500000, infinity)

```
def park_size(area):

    if ______________________________:

        ______________________________

    elif ______________________________:

        ______________________________

    else:

        ______________________________
```

(e) **(3 pt)** Now, using the `park_size` function you defined in the previous part, create a two-column table called `parks_with_sizes` that has one column containing the names of all parks as they appear in the `parks` table and another column containing the size label of each park as a string. The two columns should be named `"Name"` and `"Size"`, respectively.

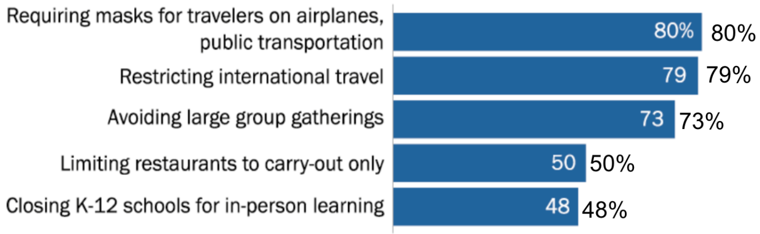
You may assume that the `park_size` function has been implemented correctly, even if you did not complete the previous part.

```
parks_with_sizes = ____________________________________________________________
```

3. (3 points) Fighting Covid

The bar chart below is from a recent PEW Research Center survey. Each bar represents the percent of U.S. adults who view the corresponding policy as necessary to address the coronavirus outbreak. The percent in each bar is provided next to it for ease of reading.

Does this bar chart display a categorical distribution? Pick the best answer from the options below. **Fill in exactly one bubble.**



○ Yes, because “Requiring masks,” “Limiting restaurants,” etc are categories.

○ No, this is a numerical distribution because percents are numbers.

○ No, this is not a distribution of any kind.

○ Maybe, maybe not. There is not enough information to decide.

4. (6 points) Mysterious Figure

A Data 8 student has defined a function called `repeat_it`. The function takes two arguments.

- The first argument is a function that takes no arguments and returns a numerical value.
- The second argument is a positive whole number.

The expression `repeat_it(function, n)` evaluates to an array of the results of `n` repetitions of calling `function`.

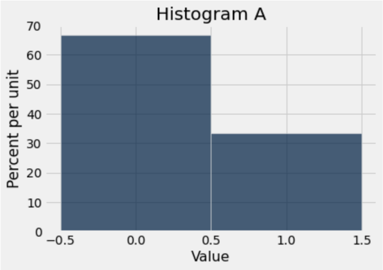
Here is an example.

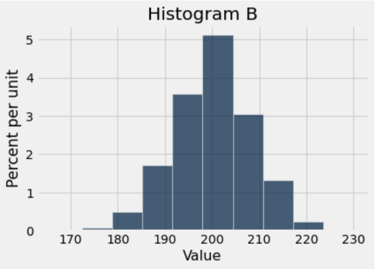
```
def example_function():
    return 10

repeat_it(example_function, 3)
```

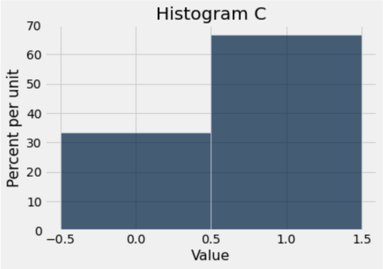
Running the example code yields the same array as you get from the call `make_array(10, 10, 10)`.

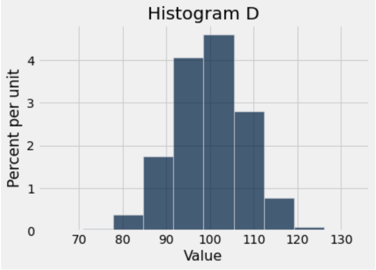
The code below produces one of the histograms (A)-(E).

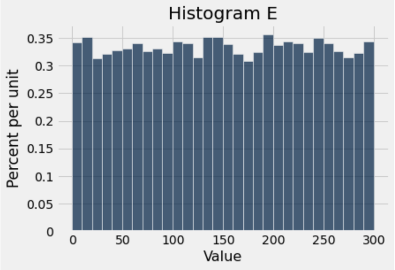
```
def my_function():
    t = Table().with_column(’Value’, make_array(0, 1, 1))
    return sum(t.sample(300).column(’Value’))

Table().with_column(
    ’Value’, repeat_it(my_function, 10000)
    ).hist()
```











(a) **(3 pt)** Which histogram is produced by the code? **Fill in exactly one bubble.**

○ A ○ B ○ C ○ D ○ E

(b) **(3 pt)** In the box below, briefly explain your choice of histogram.

5. **(6 points)** **Song Durations**

Christina is interested in learning more about the duration of songs on Spotify. She collects a random sample of 400 songs listed on the platform and stores the data in the table `songs`, which has one column labelled `"Duration"`. The average song duration in the sample is 185 seconds and the standard deviation is 25 seconds. Christina wants to use this sample of songs to make some estimates about the population of songs and their durations. ”

(a) **(3 pt)** Define a function `song_ci` that constructs a 95% confidence interval for the population mean as follows and returns it as an array. The function takes in the argument `reps`, the number of bootstrap repetitions wanted.

```
def song_ci(reps):

    stats = ____________________________________________

    for ________________________________________________:

        resample = _____________________________________

        new_mean = _____________________________________

        stats = ________________________________________

    left_end = _________________________________________

    right_end = ________________________________________

    return _____________________________________________
```

(b) **(3 pt)** Christina creates an interval by using `song_ci(10000)`. To get a more accurate estimate at the same level of confidence, Christina would like to create a new 95% confidence interval that is half as wide as this one. Which one of the following do you think is the best advice for her? **Fill in exactly one bubble.**

○ She should use `song_ci(20000)`

○ She should use a sample of size 800

○ She should use `song_ci(40000)`

○ She should use a sample of size 1600

6. (21 points) Birth Days

(a) (3 pt) Many simulations involve carrying out the same chance-based process repeatedly and generating an array of simulated values. Define a function `repeat_it` that takes two arguments:

- The first argument is a function that takes no arguments and returns a numerical value. If `f` is such a function, remember that to call it you have to use `f()`.
- The second argument is a positive whole number.

The expression `repeat_it(f, n)` evaluates to an array of the results of `n` repetitions of calling `f`.

Define the function in the box below.

(b) (3 pt) Now for the data. A doctor studying births in a large hospital system asks you to determine whether or not births in the system are distributed evenly over the week.

To help you make your decision, the doctor has gathered data on 1000 randomly sampled births in the system. Here is the distribution of days of the week for the 1000 births in his sample.

	Sunday	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
Proportion of births	0.11	0.14	0.15	0.15	0.17	0.15	0.13

In case you need it later, we have put the proportions in an array.

```
data = make_array(0.11, 0.14, 0.15, 0.15, 0.17, 0.15, 0.13)
```

From the following options, **SELECT ALL** that are correct statements of the null hypothesis.

○ The distribution of days in the sample is the same as the distribution of days in the population. Any difference is due to chance.

○ The 1000 days in the sample are drawn uniformly at random with replacement from the seven days of the week.

○ The 1000 days in the sample are drawn uniformly at random with replacement from the distribution given in the data table above.

○ Each of the 1000 babies has an 13% chance of being born on Sunday, regardless of all the other babies.

○ Each of the 1000 babies has a 1/7 chance of being born on Sunday, regardless of all the other babies.

○ Each of the 1000 babies has an equal chance of being born on any day of the week, regardless of all the other babies.

(c) **(3 pt)** Select which **one** of the following is a correct alternative hypothesis.

- ○ Babies are more likely to be born on Thursday than on any other day of the week.
- ○ The model in the null hypothesis is incorrect.
- ○ The model in the null hypothesis overestimates the proportion of births on Sundays.
- ○ The distribution of the days in the sample is different from the distribution of days in the population.

(d) **(3 pt)** Choose an appropriate test statistic to conduct this hypothesis test.

(e) **(3 pt)** Write a Python expression that evaluates to the observed test statistic.

____

(f) **(3 pt)** To carry out the hypothesis test, you must simulate your test statistic. Define a function `simulate_one` that takes no argument and returns one value of your test statistic simulated under appropriate assumptions.

```
def simulate_one():


    sim_data = ____________________________________________


    return ______________________________________________
```

(g) **(3 pt)** Suppose you decide to simulate the test statistic 5000 times and use 1% as the cutoff for the *p*-value of the test. Fill in the blank in the code below.

For the test to reject the null hypothesis, the observed test statistic has to be bigger than the value of the following expression:

```
percentile(___________________, repeat_it(simulate_one, 5000))
```

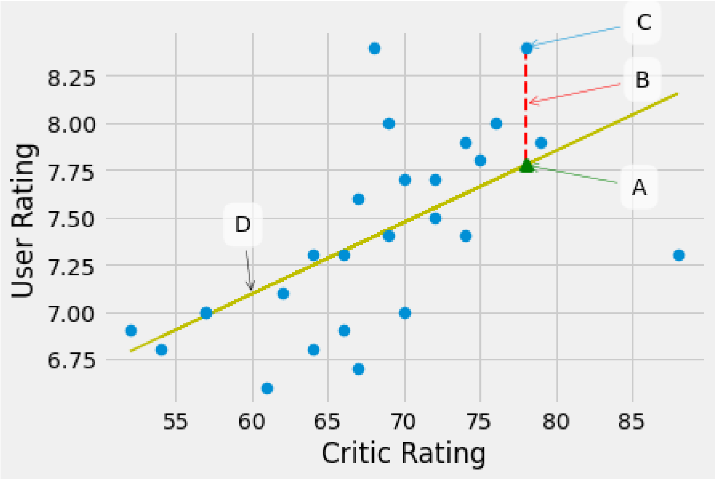
7. (12 points) Movie Reviews

Sonya is interested in seeing how IMDb user reviews differ from critic's opinions on movies. She collects data on User Rating and Critic Rating for every movie released by Marvel Studios in the past 15 years.

(a) **(3 pt)** Since critics often see the movie before regular users, Sonya is interested in predicting User Rating from Critic Rating. Which of the following techniques would help her do so? **SELECT ALL** that are correct.

- ○ Classification
- ○ Regression Equations
- ○ None of the five choices above
- ○ Bootstrapping
- ○ Simulation
- ○ Method of Least Squares

(b) (3 pt) After inspecting her data, Sonya reasons that a linear model would be a good fit, and creates one for her data. Below is her visualization with certain aspects labeled A, B, C, or D. Match each term below to its label on the graph, or “Not Pictured”. Some letters may be used multiple times or not at all.



i) Predicted Value

○ A ○ B ○ C ○ D ○ Not Pictured

ii) Residual

○ A ○ B ○ C ○ D ○ Not Pictured

iii) Line of Best Fit

○ A ○ B ○ C ○ D ○ Not Pictured

iv) Intercept

○ A ○ B ○ C ○ D ○ Not Pictured

v) RMSE

○ A ○ B ○ C ○ D ○ Not Pictured

vi) Observed Value

○ A ○ B ○ C ○ D ○ Not Pictured

vii) Fitted Value

○ A ○ B ○ C ○ D ○ Not Pictured

(c) **(3 pt)** In the graph, Sonya finds the correlation between Critic Rating and User Rating to be 0.6. Suppose she now adds a new movie to her dataset with a Critic Rating of 60 and a User Rating of 7.8. Will the correlation increase, decrease, or stay the same?

○ Increase ○ Decrease ○ Stay the same

(d) **(3 pt)** Explain your choice above.

8. **(6 points)** **Prediction Error**

Students in a class take two tests called Midterm and Final. The professor has developed a model to predict Final scores based on Midterm scores, using the data from past semesters' offerings of the course.

The table `predictions` contains a column `'Midterm'` consisting of each possible Midterm score. The only possible Midterm scores are 0, 1, 2, ..., 100, and therefore the table has 101 rows.

For each possible Midterm score, the second column `'Predicted Final'` contains the corresponding predicted Final score based on the professor's model.

Midterm	Predicted Final
0	23.17
1	23.34
2	23.52

...(98 rows omitted)

This semester's class has 200 students. After the students take both tests, the professor creates a table called `scores` that has one row for each of the 200 students. The column `'Midterm'` contains the student's score on Midterm. The column `'Final'` contains the student's score on Final.

Midterm	Final
69	39
63	47
63	45

...(197 rows omitted)

(a) **(3 pt)** Create a table `errors` that has one row for each of the 200 students, and four columns. In each row,

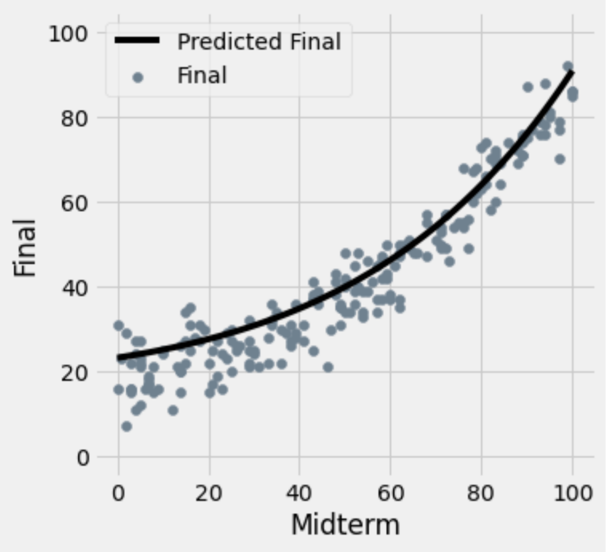
- The column `'Midterm'` should have the student's Midterm score.
- The column `'Final'` should have the student's Final score.
- The column `'Predicted Final'` should have the student's predicted Final score based on the professor's model.
- The column `'Error'` should have the difference between the student's Final score and predicted Final score.

For clarity, here is a randomly selected row of the table `errors` that you are going to create.

Midterm	Final	Predicted Final	Error
91	72	77.4	-5.4

```
t = ____________________________________________________________

e = ____________________________________________________________

errors = _______________________________________________________
```

(b) **(3 pt)** The graph below shows the actual and predicted scores. It is fine to assume that each dot represents exactly one student.



Use the scatter plot to pick the statement that evaluates to ‘True‘.

○ `np.average(errors.column('Error')) < 0`

○ `np.average(errors.column('Error')) == 0`

○ `np.average(errors.column('Error')) > 0`

9. (15 points) Ages

A data scientist takes a random sample of 400 people in a large city. The ages of the sampled people have an average of 35 years and an SD (standard deviation) of 20 years.

The data scientist bootstraps the sample 10,000 times, calculates the mean age of each bootstrapped sample, and finds the interval that contains the middle 95% of the 10,000 bootstrapped means. The interval goes from 33 years to 37 years.

(a) (3 pt) The interval (33 years, 37 years) is an approximate 95% confidence interval for the _______ of the people in the _______.

Fill in the blanks above by selecting from the following options.

(i) Blank 1 (make **exactly one** choice):

○ ages ○ average age ○ average
○ sample ○ sample mean ○ city ○ city mean

(ii) Blank 2 (make **exactly one** choice):

○ ages ○ average age ○ average
○ sample ○ sample mean ○ city ○ city mean

(b) (3 pt) The distribution of the ages of the sampled people (pick **exactly one** option):

○ is approximately normal by the Central Limit Theorem.
○ is approximately normal, but not because of the Central Limit Theorem.
○ is not normal, not even approximately.
○ may be approximately normal, or not; we need more information to decide.

(c) (3 pt) True or false: Approximately 95% of the people in the sample are between 33 and 37 years old.

○ True ○ False

(d) (3 pt) True or false: Approximately 95% of the people in the city are between 33 and 37 years old.

○ True ○ False

(e) (3 pt) The city is in a country where the average age is 35.5 years. If possible, perform a statistical test of whether or not the average age in the city is 35.5 years, using 1% as the cutoff for the p-value. State your conclusion by picking **exactly one** of the options below.

○ Since the p-value cutoff and the confidence level of the interval are inconsistent, we cannot perform this test.

○ The test concludes that the data are consistent with the hypothesis that the average age in the city is 35.5 years.

○ The test concludes that the data are not consistent with the hypothesis that the average age in the city is 35.5 years.

10. (9 points) Coffee Consumption

Meghan wants to estimate the difference between the coffee consumption of Data 8 students and Data 100 students. To help answer this question, she will take a random sample of 150 students from each of the two classes next term. You can assume that each class will have well over 1000 students and that no student will be enrolled in both classes.

Meghan will measure each sampled student’s coffee consumption by the total amount (in ounces) of coffee that the student will drink during the Spring semester. She will put the results in a table `data` that has one row for each of the 300 sampled students, one column `’Coffee’` containing the student’s coffee consumption, and one column `’Class’` containing the string `’Data 8’` or `’Data 100’` depending on which class the student is taking.

She will then create two new tables as follows:

```
data8 = data.where(’Class’, are.equal_to(’Data 8’))
data100 = data.where(’Class’, are.equal_to(’Data 100’))
```

Meghan would like to estimate the following parameter: the difference between the mean consumption of coffee in Data 8 and the mean consumption of coffee in Data 100. Define this difference as Data 8 mean - Data 100 mean.

(a) (1.5 pt) Consider the following process:

- Repeat the following 10,000 times:
  - Bootstrap the table `data8` and compute the mean of the `Coffee` column of the bootstrapped table.
  - Bootstrap the table `data100` and compute the mean of the `Coffee` column of the bootstrapped table.
  - Find the difference between the bootstrapped Data 8 mean and the bootstrapped Data 100 mean.
- Find the endpoints of the interval created by the middle 99% of the 10,000 differences.

Is that a correct way of creating an approximate 99% confidence interval for the parameter?

○ Yes ○ No

(b) (1.5 pt) If you chose “No” above, why not? Briefly explain in the box below. If you chose “Yes,” you don’t have to write anything.

(c) **(1.5 pt)** Is the following process a correct way of creating an approximate 99% confidence interval for the parameter?

- Repeat the following 10,000 times:
    - Shuffle the rows of `data8` and compute the mean of the `Coffee` column of the shuffled table.
    - Shuffle the rows of `data100` and compute the mean of the `Coffee` column of the shuffled table.
    - Find the difference between the `Coffee` mean of the shuffled `data8` table and the `Coffee` mean of the shuffled `data100` table.
- Find the endpoints of the interval created by the middle 99% of the 10,000 differences.

○ Yes ○ No

(d) **(1.5 pt)** If you chose "No" above, why not? Briefly explain in the box below. If you chose "Yes," you don't have to write anything.

(e) **(1.5 pt)** Is the following process a correct way of creating an approximate 99% confidence interval for the parameter?

- Repeat the following 10,000 times:
    - Shuffle the `Class` column of `data` and create a new table by attaching the shuffled class labels to the original column `Coffee` of the `data` table.
    - Group this table by the shuffled labels and find the difference between the `Data 8` group mean and the `Data 100` group mean.
- Find the endpoints of the interval created by the middle 99% of the 10,000 differences.

○ Yes ○ No

(f) **(1.5 pt)** If you chose "No" above, why not? Briefly explain in the box below. If you chose "Yes," you don't have to write anything.

11. (3 points) Positive Test

Doctors in a city have access to a medical test for a disease that affects 1% of the people in the city. The test has high accuracy:

- For a person who has the disease, the test returns a positive result with chance 98%.
- For a person who does not have the disease, the test returns a negative result with chance 99%.

A person in the city has symptoms of the disease and visits their local doctor. The doctor examines the patient and recommends that the patient take the test. The test result comes back positive, and the doctor turns to you for advice, asking, "Now that we know the test result is positive, what is the chance that the person has the disease?" Which of the following is your answer? **Fill in exactly one bubble.**

○ 0.01 ○ 0.98 ○ 0.01×0.98

○ $\dfrac{0.01 \times 0.98}{(0.01 \times 0.98) + (0.99 \times 0.01)}$ ○ $\dfrac{0.01 \times 0.98}{(0.01 \times 0.98) + (0.99 \times 0.99)}$

○ I went through all the calculations above and none of them is valid.

12. (12 points) Movie Directors

In each part of this question, you are free to use tables and functions that have been defined earlier in the question, even if you couldn't define them correctly.

(a) (3 pt) The Python function `np.unique` takes an array as its argument and returns an array consisting of the distinct elements of the argument array. Here is an example of its use.

```
example_array = make_array(’cat’, ’cat’, ’dog’, ’bear’, ’bear’, ’dog’, ’tiger’, ’bear’)
np.unique(example_array)

>>> array([’bear’, ’cat’, ’dog’, ’tiger’], dtype=’<U5’)
```

The output is an array consisting of the four distinct elements in `example_array`.

Define a function `count_distinct` that takes an array as its argument and returns the number of distinct elements in the array. Provide your definition in the box.

(b) **(3 pt)** Each row of the table `directors` corresponds to a movie released by a major Hollywood studio in the years 1980 through 2019. The table has five columns.

• `'Movie'` contains the name of the movie.

• `'Studio'` contains the name of the studio that released the movie.

• `'Year'` contains the year in which the movie was released.

• `'Decade'` contains the decade in which the movie was released. There are four decades: '1980' consists of the years 1980 through 1989, '1990' consists of the years 1990 through 1999, and so on.

• `'Director'` contains the name of the director of the movie. You can assume that only one director is listed for each movie.

The table has numerous rows. To show you what it looks like, here are just three of the rows in which the director is J.J. Abrams. He has directed many movies and appears in other rows as well.

Movie	Studio	Year	Decade	Director
Mission Impossible	Paramount	2006	2000	J.J. Abrams
Super 8	Paramount	2011	2010	J.J. Abrams
Star Wars: The Force Awakens	Disney	2015	2010	J.J. Abrams

Complete the code below so that the last line evaluates to a table consisting of two columns:

• The first column should be labeled `'Studio'` and contain all the distinct studios.

• The second column should contain the number of different directors whose movies were released by the studio in the years 1980 through 2019.

```
t1 = ____________________________________________________________
t1
```

(c) **(3 pt)** Complete the code below so that the last line evaluates to a table that has five columns:

• A column containing all the distinct studios

• A column for each of the four decades

For each studio, the values in each decade column should contain the number of different directors whose movies were released by the studio in that decade.

```
t2 = ____________________________________________________________
t2
```

(d) **(3 pt)** Which of the following does the expression `t1.column(1) - t2.drop(0).apply(sum)` evaluate to? Pick **exactly one** choice.

(Technical note: You do not need to worry about numerical inaccuracy or roundoff: all numbers are `int`s, so that won't happen.)

○ An array in which all the values are 0

○ An array in which some of the values are not 0

○ The expression generates an error message.

13. (6 points) Animated Movies

Professor Wagner wants to build a classifier to predict whether a movie is animated or not, based on the script. He hires an intern to download 64,000 movie scripts and identify which are animated. He randomly shuffles this dataset and then splits it into a training set with 32,000 movies and a test set with 32,000 movies, builds a 1-nearest neighbor classifier (with k=1) using the training set, and then measures its accuracy on the test set. His classifier gets 95% accuracy on the test set.

The next day, Prof. Wagner realizes that the intern took a shortcut. The intern built the dataset by downloading 16,000 different random movies and making 4 identical copies of each movie.

(a) (3 pt) Which of the following do you think is most likely to be true?

○ The accuracy on a randomly selected new movie will be significantly less than 95%.
○ The accuracy on a randomly selected new movie will be about 95%.
○ The accuracy on a randomly selected new movie will be significantly more than 95%.

(b) (3 pt) In the box below, briefly explain the reasoning behind your choice.

14. (0 points) Last Words (optional)

If there was any question on the exam that you thought required clarification to be answerable, please identify the question and state the assumptions you made in your answer. Please note that we will only consider this information if we agree that the question required clarification and that your assumptions were reasonable.

15. (0 points) Write your name in the space provided on one side of every page of the paper exam. You're done!