Data 8 Spring 2019 — Foundations of Data Science

Final Exam

INSTRUCTIONS

- The exam is worth 150 points. You have 170 minutes to complete it.
- The exam is closed book, closed notes, closed computer/phone/tablet, closed calculator, except the official final exam reference guide provided with the exam.
- Write/mark your answers on the exam in the blanks/bubbles 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.
- If you need scratch paper, you are welcome to use the reference guide and the back of this cover page. Scratch work 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.

Last name	Solutions
First name	
Student ID number	
Calcentral email (`_@berkeley.edu`)	
Lab GSI	
Your seat number (e.g. A1) & room	
← 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)**	

This page is intentionally left blank. You can use it for scratch work but it will not be graded.

1. (10 points) Python expressions

For each expression below, say what it evaluates to.

(a) (2 pt) `np.count_nonzero(make_array(0,1,0,1,0))`

○ 3 ● 2 ○ 5 ○ Python returns an Error

(b) (2 pt) `np.diff(make_array(1,2,3,4))`

○ [1,2,3] ○ [-1,-1,-1] ● [1,1,1] ○ Python returns an Error

(c) (2 pt) `make_array(1,2,3) + np.arange(2,8,3)`

○ [3,10,6] ○ [3,7,11] ○ [3,10] ● Python returns an Error

(d) (2 pt) `percentile(10, make_array(1,10,20,100))`

● 1 ○ 10 ○ 20 ○ Python returns an Error

(e) (2 pt) `np.std(make_array(0,10))`

○ 4 ● 5 ○ 6 ○ Python returns an Error

2. (8 points) Basketball

A basketball player is attempting shots from a fixed distance away from the basket. Assume that each shot is good (that is, goes in the basket) with chance 6/10 independently of all other shots.

(a) (3 pt) If the player attempts two shots, what is the chance that both the shots are good?

○ 12/100 ○ 30/90 ● 36/100 ○ 8/100
○ Cannot be calculated with the information given

(b) (5 pt) Pick **ALL** that are true:

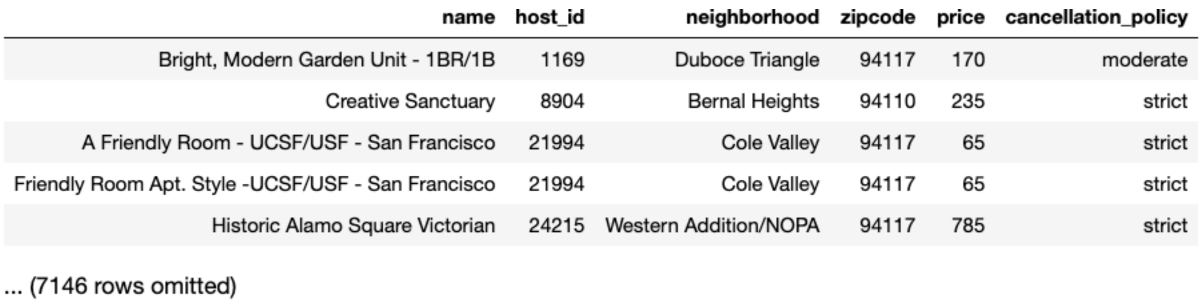
The player will take 200 shots in the next practice. The proportion of good shots

○ has chance 40% of being 0 and chance 60% of being 1
● has many possible values and the histogram of its probability distribution looks like the normal curve
● has a probability distribution whose histogram balances at 0.6

3. (26 points) Airbnb

A table `airbnb` stores information about all of the Airbnb listings in San Francisco. It has one row for each listing and six columns:

- **name**: a string, the name of the listing
- **host_id**: an int, the numerical ID of the host
- **neighborhood**: a string, the neighborhood where the listing is located
- **zipcode**: an int, the ZIP code where the listing is located
- **price**: a float, the price in dollars for a one-night stay
- **cancellation_policy**: a string, the cancellation policy, which is `'flexible'`, `'moderate'`, or `'strict'`



name	host_id	neighborhood	zipcode	price	cancellation_policy
Bright, Modern Garden Unit - 1BR/1B	1169	Duboce Triangle	94117	170	moderate
Creative Sanctuary	8904	Bernal Heights	94110	235	strict
A Friendly Room - UCSF/USF - San Francisco	21994	Cole Valley	94117	65	strict
Friendly Room Apt. Style -UCSF/USF - San Francisco	21994	Cole Valley	94117	65	strict
Historic Alamo Square Victorian	24215	Western Addition/NOPA	94117	785	strict

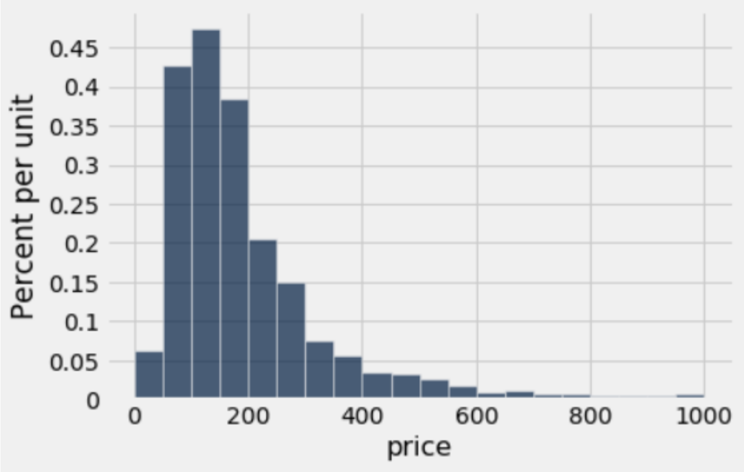
... (7146 rows omitted)

In some parts below, you have to fill in the blanks in Python expressions. **You must use ONLY the lines provided.** Some of the chained operations we might normally do in one line have been broken up into two or more lines, storing intermediate results in temporary tables or arrays. Do not write any code outside the blanks provided. The expression in the last line should evaluate to the value described.

(a) **(5 pt)** Fill in the blanks below to find the ZIP code with the highest total number of listings.

```
by_zipcode = airbnb.____group____(____‘zipcode’____)

in_order = by_zipcode.____sort____('count', ____descending = True____)

in_order.____column(‘zipcode’).item(0)____
```

(b) **(3 pt)** The plot below shows the distribution of the listing prices in the data set. The bins all have width of $50 (the plot excludes a very small number of listings with price greater than $1000).



Choose the correct line of code to create the plot.

○ `airbnb.bar(’price’, bins = np.arange(0,1001,50))`
● `airbnb.hist(’price’, bins = np.arange(0,1001,50))`
○ `airbnb.hist(’price’, bins = np.arange(1000))`
○ `airbnb.group(’price’, bins = np.arange(0,1001,50)).bar(’count’)`

(c) (3 pt) Based on the plot in part (b), which of the following choices is closest to the percent of listings that have prices less than $200?

○ 38% ○ 0.38% ○ 19% ● 66% ○ 50%
○ This cannot be determined from the plot

(d) (2 pt) Based on the plot shown above, what would you say about the relationship between the mean price and the median price?

○ The median price is larger than the mean price.
● The mean price is larger than the median price.
○ We cannot tell from the plot whether the mean price is larger than the median price.

(e) (6 pt) Next we will investigate which hosts operate more than one listing (we’ll call these hosts “serial hosts”). Complete the code below to create a table called `hosts` with **one row for each host** and three columns:

- **host_id**, an int, the host’s numerical ID
- **num_listings**, an int, the total number of San Francisco listings operated by the host
- **serial_host**, a boolean, `True` if `num_listings` is 2 or more, and `False` otherwise

`hosts = airbnb.` group `(` ‘host_id’ `)`

`hosts.relabeled(` ‘count’ `, ’num_listings’)`

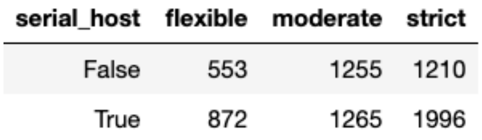
`serial_host_array =` hosts.column(‘num_listings’) `>= 2`

`hosts = hosts.` with_column `(` ‘serial_host’ `,` serial_host_array `)`

(f) (4 pt) Create a new table called `airbnb2` that has **one row for each listing**, all the columns of `airbnb`, a new column `num_listings` indicating how many total listings the listing’s host operates, and a new column `serial_host` indicating whether the listing’s host is a serial host.

`airbnb2 = airbnb.` join `(` ‘host_id’ `,` hosts `)`

(g) (3 pt) Finally, make a table showing the number of listings with each cancellation policy, for each host type (serial or not serial). Your code should evaluate to the table shown below:



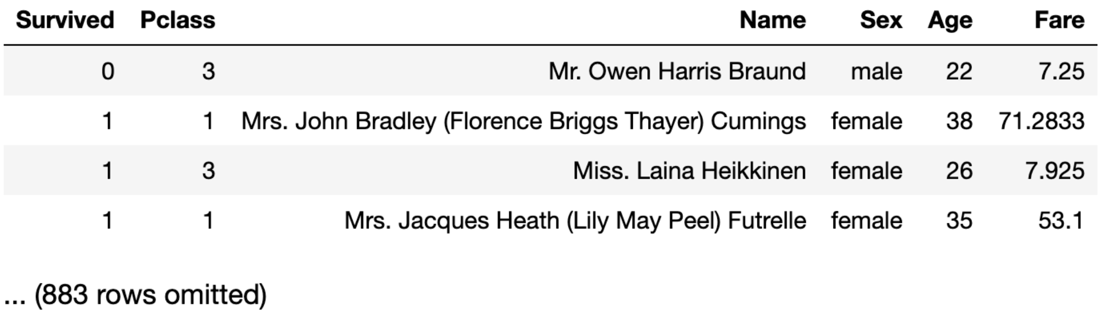
serial_host	flexible	moderate	strict
False	553	1255	1210
True	872	1265	1996

`airbnb2.` pivot(‘cancellation_policy’, ‘serial_host’)

4. (15 points) The Titanic

A table `titanic`, shown below, stores information about the passengers on the Titanic when it sank in 1912. You may assume the data set includes all of the passengers who were on the ship. It has one row for each passenger and six columns:

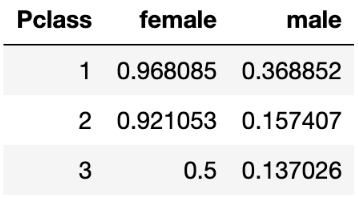
- **Survived**: an int, 1 if the passenger survived and 0 otherwise
- **Pclass**: an int, 1, 2, or 3, denoting that the passenger was traveling in first, second, or third class
- **Name**: a string, the name of the passenger
- **Sex**: a string, coded as `’male’` or `’female’`
- **Age**: an int, the age in years of the passenger
- **Fare**: a float, the fare in pounds that the passenger paid



Survived	Pclass	Name	Sex	Age	Fare
0	3	Mr. Owen Harris Braund	male	22	7.25
1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings	female	38	71.2833
1	3	Miss. Laina Heikkinen	female	26	7.925
1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35	53.1

... (883 rows omitted)

(a) **(4 pt)** The `’Sex’` and `’Pclass’` columns are important predictors of whether passengers survived, because the crew gave priority to women, children, and upper-class passengers when boarding the lifeboats. Fill in the blank in the code to produce the table below, which gives the proportion of male and female passengers in first, second, and third class who survived. For example, among the female passengers in third class exactly 50% survived.

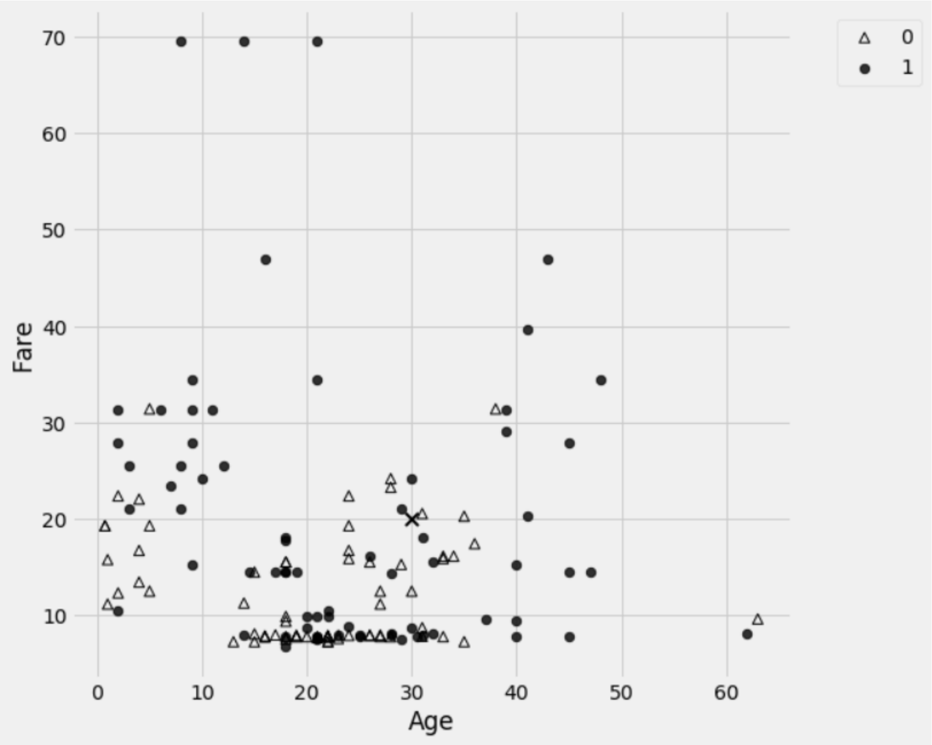


Pclass	female	male
1	0.968085	0.368852
2	0.921053	0.157407
3	0.5	0.137026

`titanic.` pivot (‘Sex’, ‘PClass’, ‘Survived’, np.mean)

(b) **(3 pt)** The scatter plot (on the next page) contains one point for each female passenger in third class. The plot shows the `Age` and `Fare` for each passenger, with `Survived` represented by the shape of the point. The triangles correspond to passengers who died (0) and the circles correspond to passengers who survived (1).

Suppose we train a k-nearest neighbor classifier to predict whether a female third-class passenger survived given her age and the fare she paid. We will use our classifier to predict what will happen to a hypothetical 30-year-old woman who paid 20 pounds (marked by the symbol × on the scatter plot).



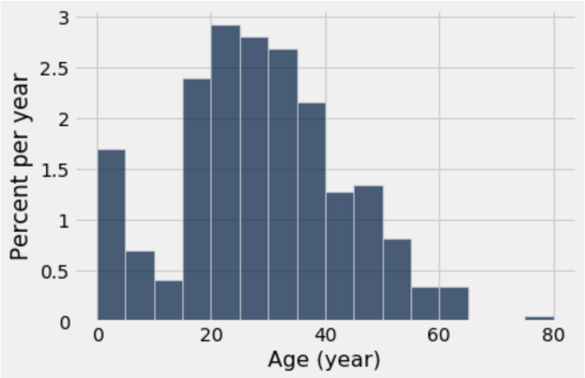
If we use a 3-nearest-neighbor classifier, what will we predict for the woman marked by the symbol ×?

● Survive ○ Not survive ○ The prediction cannot be determined from the plot

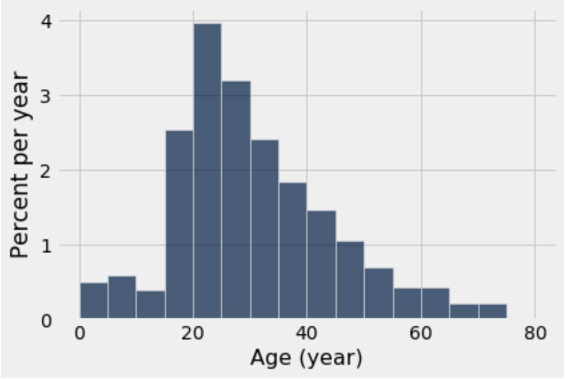
(c) **(3 pt)** If we use a 13-nearest-neighbor classifier, what will we predict for the same woman?

○ Survive ● Not survive ○ The prediction cannot be determined from the plot

(d) **(5 pt)** The histograms below show the age distributions of passengers who survived (left plot) and passengers who did not survive (right plot).



Age distribution of survivors



Age distribution of non-survivors

Which of the following statements can we justify based on what we can see in the two above charts? Select **ALL** answers that apply.

○ Most of the passengers under age 5 survived
● Most of the survivors were aged 20 or older
○ The crew placed a higher priority on saving young children than it did on saving adults
● None of the passengers between the ages of 65 and 74 survived

5. (10 points) Death Penalty

Each year, a survey organization takes a random sample of U.S. adults and asks the sampled adults whether they favor the death penalty for murder. Among the adults sampled in 2016, 49% answered yes to the question. Among the adults sampled in 2018, 54% answered yes to the question.

The survey organization wants to test whether the percent of U.S. adults who would answer yes to the question increased between 2016 and 2018.

(a) (5 pt) Help the survey organization perform a test of its hypotheses. Pick an appropriate null hypothesis and alternative hypothesis from the **Hypotheses** options below; pick the observed value of an appropriate test statistic from the **Observed Test Statistic** options below; and pick the appropriate way to find the P-value from the **P-value** options below. Note that more than one combination might be correct; just pick any correct combination.

Hypotheses

A. In 2016, 49% of US adults would answer yes and in 2018, 54% of US adults would answer yes.
B. The percents of US adults who would answer yes are different in the two years.
C. The percents of US adults who would answer yes are the same in the two years.
D. The percent of US adults who would answer yes is greater in 2016 than in 2018.
E. The percent of US adults who would answer yes is greater in 2018 than in 2016.
F. The percent of US adults who would answer yes is 49% in both years; the 2018 sample is off due to chance.
G. The percent of US adults who would answer yes is 54% in both years; the 2016 sample is off due to chance.
H. The percent of US adults who would answer yes in 2018 is greater than that in 2016, due to chance.
I. The percent of US adults who would answer yes is 50% in both years; the samples are off due to chance.

Observed Test Statistic

A. 0.49
B. 0.54
C. -0.01
D. 0.04
E. $0.49 - 0.54$
F. $(|0.49 - 0.5| + |0.54 - 0.5|)/2$

P-Value

A. Simulate repeatedly from the distribution of the test statistic under the null hypothesis and find the proportion of simulated statistics equal to or greater than the observed statistic.
B. Simulate repeatedly from the distribution of the test statistic under the null hypothesis and find the proportion of simulated statistics equal to or less than the observed statistic.

Null Hypothesis: ○A ○B ●C ○D ○E ○F ○G ○H ○I

Alternative Hypothesis: ○A ○B ○C ○D ●E ○F ○G ○H ○I

Test Statistic: ○A ○B ○C ○D ●E ○F **P-value:** ○A ●B

(b) (2 pt) Suppose the P-value of the test comes out to be 1.5%. If the survey organization is using the 1% cutoff for the P-value, which hypothesis would the organization pick?

● Null ○ Alternative

(c) (3 pt) Suppose the P-value of the test comes out to be 1.5%. Pick **ALL** correct options from the list below.

○ There is 1.5% chance that the null hypothesis is true.
○ There is 98.5% chance that the null hypothesis is true.
○ There is 1.5% chance that the alternative hypothesis is true.
○ There is 98.5% chance that the alternative hypothesis is true.
● None of the above options is correct.

6. (8 points) Buttons

We have 1,000 bags of Data Great buttons. Each bag contains 100 buttons. In one of the bags, all 100 buttons are blue. Each of the 999 other bags contains four colors of buttons: 40 blue, 30 red, 20 green, and 10 purple.

(a) (4 pt) Suppose we pick one of the bags at random and then pick one button at random from the bag. Given that the button picked is blue, what is the chance that we picked a bag that contains other colors of buttons as well?

○ 0.999 ○ 0.999×0.4 ○ $0.001 + (0.999 \times 0.4)$

○ $\frac{0.999}{0.001 + 0.999}$ ● $\frac{0.999 \times 0.4}{0.001 + (0.999 \times 0.4)}$ ○ $\frac{0.001}{0.001 + (0.999 \times 0.4)}$ ○ $\frac{0.4}{1 + 0.4}$

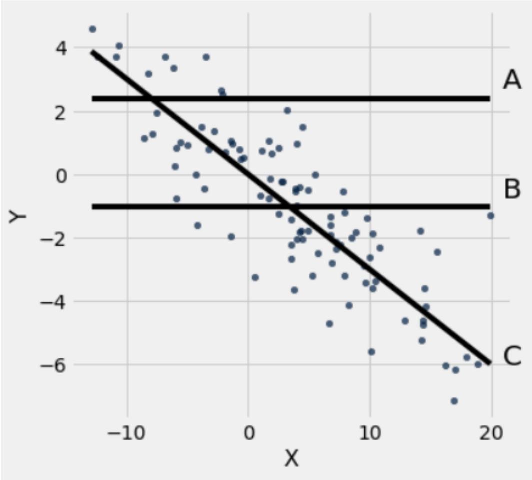
(b) (4 pt) Suppose we pick one of the bags at random and then pick two buttons at random with replacement from the bag. What is the chance that both of the buttons are blue?

○ 0.001 ○ $(0.4)^2$ ○ $0.001 + (0.4)^2$

● $0.001 + (0.999 \times (0.4)^2)$ ○ $\frac{0.001}{0.001 + (0.4)^2}$ ○ $\frac{(0.4)^2}{0.001 + (0.4)^2}$

7. (4 points) Ranking Errors

A research group is attempting to use a straight line to predict values of Y based on values of X. The scatter diagram is shown to the right along with three possible prediction lines, labeled A, B, and C:



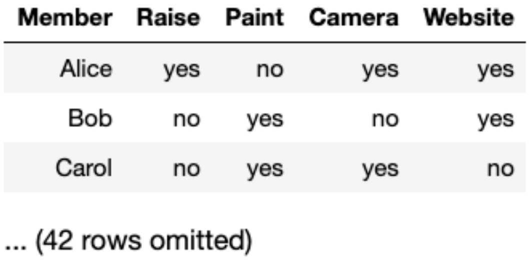
All that we know about the lines is what you can see in the plot. No further information about their equations is available. Rank the three lines in **descending order** of root mean squared error.

● A,B,C ○ A,C,B ○ B,A,C
○ B,C,A ○ C,A,B ○ C,B,A
○ This cannot be determined from the information given.

8. (15 points) Co-op Voting

A Cooperative Society (co-op) has 45 members. Each member votes yes or no on four proposals: raising parking fees, painting the hallways, installing security cameras, and revising the co-op's website.

The votes are recorded in the table `votes`, shown below. Each row corresponds to a member and the columns contains strings corresponding to that member's vote on each of the four issues. You may assume that every member did vote yes or no on all four issues.



Member	Raise	Paint	Camera	Website
Alice	yes	no	yes	yes
Bob	no	yes	no	yes
Carol	no	yes	yes	no

... (42 rows omitted)

(a) **(4 pt)** Fill in the blanks in the code below to find the proportion of members who voted yes on **both** painting the hallways and installing security cameras. The last line should evaluate to this proportion.

```
step_1 = votes.where ('Paint', are.equal_to(‘yes’))

step_2 = step_1.where ( are.equal_to(‘yes’))

step_2.num_rows / votes.num_rows
```

(b) **(3 pt)** Twenty-six out of the 45 members voted yes on painting the hallways, and 14 out of the 45 members voted yes on installing security cameras. Pick one option: The proportion of members who voted yes on **both** painting the hallways and installing security cameras is

○ $\frac{26}{45} + \frac{14}{45}$ ○ $\frac{26}{45} \times \frac{14}{45}$ ● Cannot be calculated with the information given

(c) **(4 pt)** Fill in the blanks in the code below to find the majority position on raising parking fees. The second line should evaluate to the string `'yes'` if the majority of members voted yes on raising fees, and the string `'no'` if the majority voted no on raising fees.

```
new_tbl = votes.group ('Raise').sort (‘count’, descending = True)

new_tbl.column (‘Raise’).item(0)
```

(d) **(4 pt)** Pick one option: The expression `votes.sort('Website').column('Website').item(22)` evaluates to

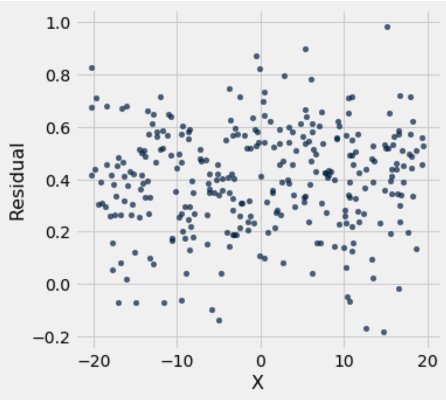
● the majority position on revising the website
○ the minority position on revising the website
○ `'yes'` or `'no'`, but we can't tell if that is the majority or minority position on revising the website

9. (8 points) Residuals

Each of the following plots represents the residuals from an attempted linear regression of a variable Y on a variable X (that is, the regression line is meant to predict values of Y based on values of X). For each one, indicate whether the regression line seems to be a good fit, or seems to be a bad fit, or if it is impossible for a residual plot to look like the plot shown.

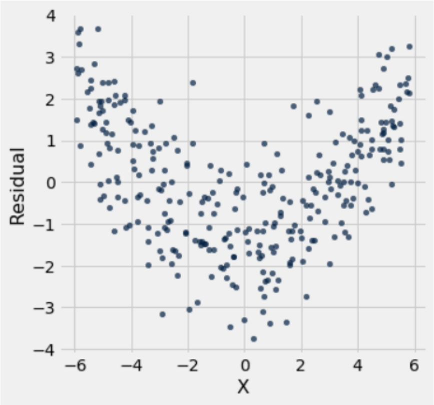
In each part, choose the best option based on what you see in the plot and just rough mental math if needed. Don't attempt any precise calculations.

(a) (2 pt)



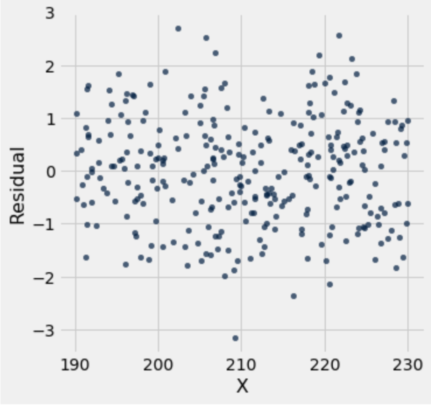
● The regression seems to fit the data well
○ The regression seems not to fit the data
○ A residual plot could never look like this

(b) (2 pt)



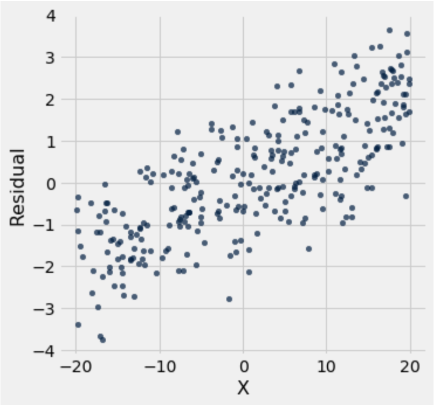
○ The regression seems to fit the data well
● The regression seems not to fit the data
○ A residual plot could never look like this

(c) (2 pt)



● The regression seems to fit the data well
○ The regression seems not to fit the data
○ A residual plot could never look like this

(d) (2 pt)



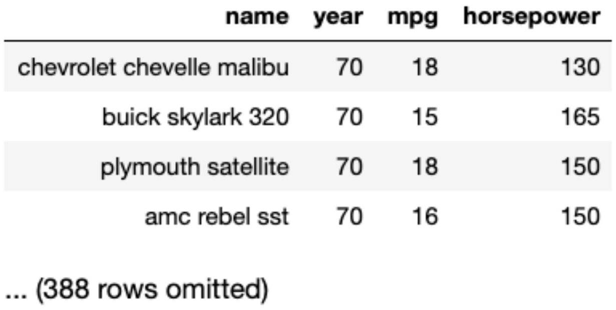
○ The regression seems to fit the data well
○ The regression seems not to fit the data
● A residual plot could never look like this

10. (24 points) Autos

A data set `auto` stores information about 392 automobile models including various performance specifications. You may assume the data set is a simple random sample of autos from a much larger population. It has one row for each auto and contains four columns:

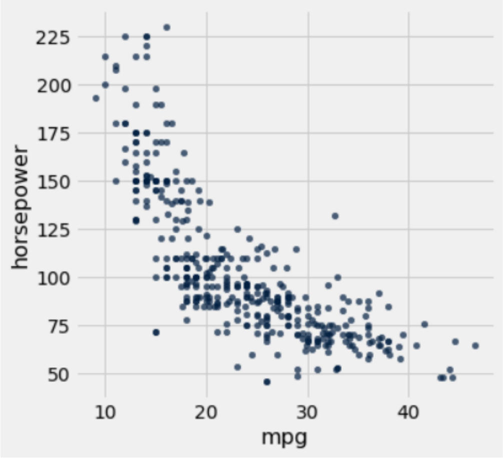
- **name**: a string, the make and model of the auto
- **year**: an int, the last two digits of the year in which the auto was made
- **mpg**: a float, the fuel efficiency of the auto in miles per gallon
- **horsepower**: a float, a measure of how powerful the engine is

There is a noticeable negative association between horsepower and mpg. Their scatter plot is shown below alongside the first few rows of the table:



name	year	mpg	horsepower
chevrolet chevelle malibu	70	18	130
buick skylark 320	70	15	165
plymouth satellite	70	18	150
amc rebel sst	70	16	150

... (388 rows omitted)



(a) (2 pt) The column `'mpg'` has mean 23 and standard deviation 7.8. The column `'horsepower'` has mean 104 and standard deviation 38. The two columns have correlation -0.8.

Suppose a data analyst forgets what we know in Data 8 about not fitting linear regressions to nonlinear scatter plots, and uses a linear regression to estimate an equation of the form:

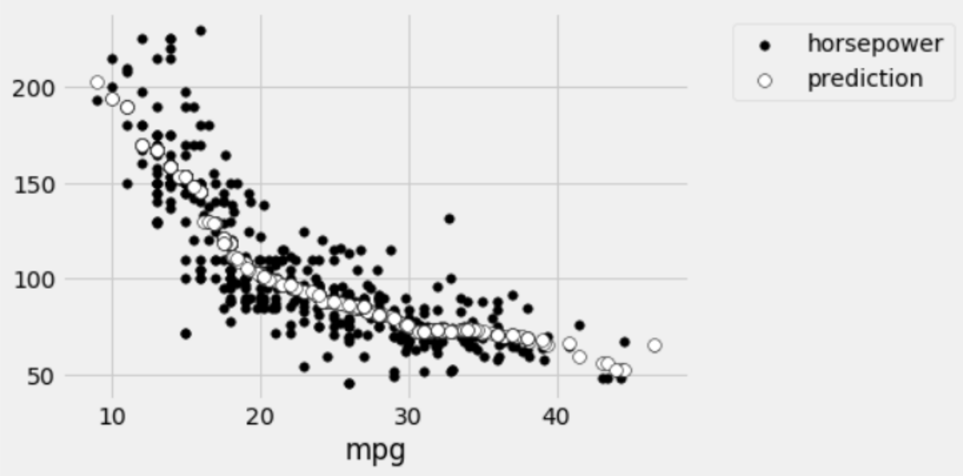
$$\text{predicted horsepower} \;=\; \text{slope} \times \text{mpg} \;+\; \text{intercept}$$

Assume the equation is in the original units of horsepower and mpg (**NOT** standard units). What is the slope of the regression line?

○ $-0.8 \times \frac{7.8}{38}$ ● $-0.8 \times \frac{38}{7.8}$ ○ $0.8 \times \frac{\sqrt{7.8}}{\sqrt{38}}$ ○ $-0.8 \times \frac{23}{104}$ ○ $0.8 \times \frac{104}{23}$

○ There is no way to know because the relationship is clearly nonlinear

(b) **(8 pt)** After looking at the scatter plot of the data, we know that the data analyst shouldn't have used a linear regression. We decide to instead plot a graph of averages, to make predictions using the simpler method of just averaging the horsepowers of all autos with similar mpg's. Define "similar" as autos whose mpg's are within 2 units of the mpg value for which we are trying to make a prediction. We get the following plot:



Complete the code below to compute the graph of averages and produce the plot above. In your code you do not need to worry about whether "similar" autos include or exclude autos that are exactly 2 mpg away. In addition, do not worry about specifying the colors or shapes of the points in the plot.

```
def predict_horsepower(x):
    """Predict the horsepower of an auto with mpg = x.
    The prediction is the average horsepower of the autos whose mpg
    is within 2 units of x"""


    close_points_tbl = auto.___where___(__‘mpg’__, __are.between(x-2 , x+2)__)


    return np.mean(close_points_tbl.__column(‘horsepower’)__________)


horsepower_preds = auto.___apply___(___predict_horsepower___, __‘mpg’__)

auto_with_preds = auto.select('mpg','horsepower').with_column('prediction', horsepower_preds)

auto_with_preds.___scatter___(___‘mpg’___)
```

(c) (5 pt) Which of the following statements are true of the linear regression the data analyst carried out in part (a) of this question? Select **ALL** statements that are definitely true. If you do not have enough information to evaluate whether the statement is true or false, **DO NOT** select it.

- ● The regression line minimizes the RMSE in this data set among all straight prediction lines
- ● The regression line minimizes the MSE in this data set among all straight prediction lines
- ○ The regression line will give essentially the same predictions as the graph of averages
- ○ The residual plot of the linear regression will have no clearly visible patterns in it
- ● The regression line has a smaller RMSE than the predictions shown in the plot in (b)

(d) (4 pt) Now suppose that we decide to try fitting a quadratic relationship to the data instead, using numerical optimization. That is, we estimate an equation of the form

$$\text{predicted horsepower} \;=\; a \times \text{mpg}^2 \;+\; b \times \text{mpg} \;+\; c$$

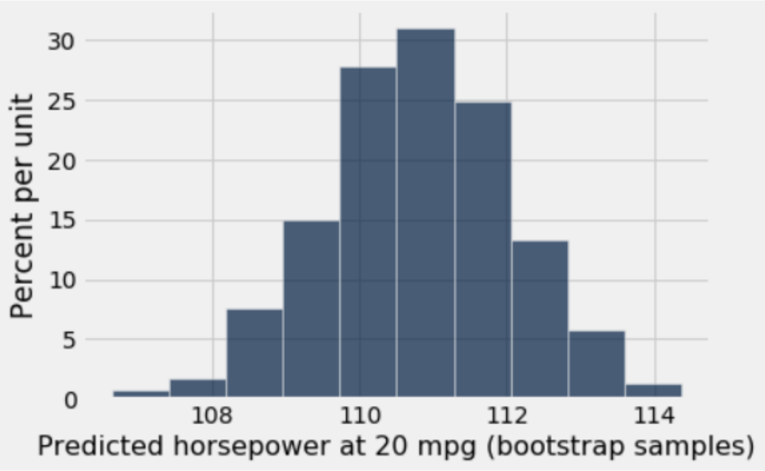
Fill in the code below to define a function that computes the RMSE for a quadratic prediction equation with values a, b, and c.

```
def quadratic_rmse(a, b, c):
    """Compute RMSE for a quadratic prediction equation of the form:
        a * (mpg ** 2) + b * mpg + c"""


    mpg = auto.column('mpg')


    hp = auto.column('horsepower')


    preds_array = a * (mpg **2 ) + b * mpg + c


    return (np.mean((hp - preds_array) ** 2)) ** 0.5
```

(e) (3 pt) Select the line of code that will use numerical optimization to find the best-fitting quadratic equation, in terms of RMSE. The code should evaluate to an array of length three whose values are the estimates of a, b, and c defining the best quadratic equation.

- ○ `minimize(a, b, c)`
- ○ `minimize(quadratic_rmse(a, b, c))`
- ○ `maximize(a, b, c)`
- ● `minimize(quadratic_rmse)`
- ○ `quadratic_rmse(minimize)`
- ○ `quadratic_rmse(auto.minimize())`
- ○ `auto.minimize(quadratic_rmse(a,b,c))`

(f) (2 pt) The plot below shows the distribution of the quadratic regression estimate for the predicted horsepower at 20 mpg in 1000 bootstrap samples:



The 5th percentile is 108.6 and the 95th percentile is 112.9. You may assume that the true relationship between the two variables is quadratic: that is, for some true values a, b, and c the horsepower for a car is the true quadratic curve $a \times \text{mpg}^2 + b \times \text{mpg} + c$, plus independent normally distributed noise with mean zero.

Select **ALL** statements that are correct. If there is not enough information to evaluate whether the statement is true or false, **DO NOT** select it.

○ If we sample 1000 more autos with fuel efficiency of 20 mpg from the same population, then about 900 of them will have horsepower between 108.6 and 112.9.

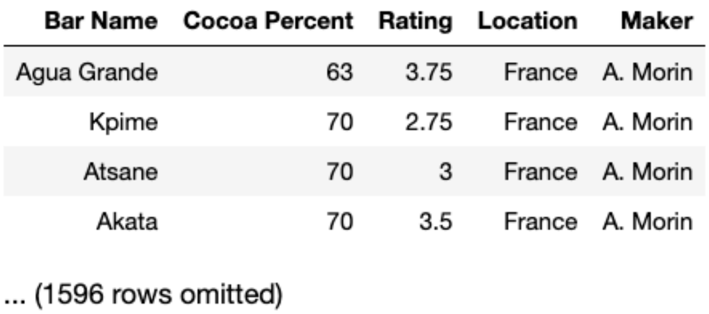
○ There is about a 90% chance that the true quadratic curve at mpg = 20 is between 108.6 and 112.9.

● If we compute a 95% confidence interval for the true quadratic curve at mpg = 20, using the same data and same bootstrap samples, it will include both 108.6 and 112.9.

○ None of the above statements is correct.

11. (22 points) Chocolate

A data set stores information about 1,600 chocolate bars. The table `chocolate` has one row for each bar:



Bar Name	Cocoa Percent	Rating	Location	Maker
Agua Grande	63	3.75	France	A. Morin
Kpime	70	2.75	France	A. Morin
Atsane	70	3	France	A. Morin
Akata	70	3.5	France	A. Morin

... (1596 rows omitted)

The table contains five columns:

- **Bar Name**: a string, the name of the chocolate bar
- **Cocoa Percent**: a float, the cocoa content of the chocolate bar by percentage (higher means darker chocolate)
- **Rating**: a float, the average rating on a scale of 1 (lowest) to 5 (highest) by a panel of expert raters
- **Location**: a string, the name of the country where the chocolate bar is made
- **Maker**: a string, the company that manufactures the chocolate bar

You may assume that this is a simple random sample from a much larger population of chocolate bars.

(a) (3 pt) In our sample, the mean Cocoa Percent is 72 and the standard deviation is 6. Pick **ALL** the intervals that we can be sure contain at least 75% of the data in the sample.

○ $72 \pm \frac{6}{1600}$ ○ $72 \pm \frac{12}{1600}$ ○ $72 \pm \frac{6}{40}$ ○ $72 \pm \frac{12}{40}$ ○ 72 ± 6 ● 72 ± 12

○ There is no way to know for sure without more information

(b) (4 pt) Belgium and Switzerland are two European countries known for their chocolate. Suppose we want to know which country's chocolate bars are rated higher on average in the population. In our sample, Belgium has mean rating 3.09 while Switzerland has mean rating 3.34, giving a mean difference of 0.25 points in the sample, which we can use as an estimate of the mean difference in the population. We will use the bootstrap to help us quantify the uncertainty of this estimate.

Complete the code below to write a function `ave_rating` that takes in a table `tbl` with the same column labels as `chocolate`, and a location `loc`, and computes the mean rating for all chocolate bars whose `Location` value is equal to `loc`. For example, `ave_rating(chocolate, 'Belgium')` should return 3.09.

```
def ave_rating(tbl, loc):
        """Compute the average rating of bars in the table tbl with Location loc."""

        loc_tbl = tbl.where(_'Location', are.equal_to('loc')_______________________)

        return _np.mean(loc_tbl.column('Rating')________________________________
```

(c) **(7 pt)** Next, complete the code below to generate 5,000 bootstrap samples, compute the mean difference between the ratings of the Swiss and Belgian chocolate bars (that is, Swiss average minus Belgian average) in each bootstrap sample, and store all of the results in the array `boot_diffs`. You may assume that `ave_rating` has been defined correctly.

```
boot_diffs = make_array()

for i in np.arange(5000):

        boot_table = chocolate.sample()

        ave_switz = ave_rating(boot_table ,’Switzerland’)

        ave_belgium = ave_rating(boot_table ,’Belgium’)

        boot_diffs = np.append(boot_diffs, ave_switz - ave_belgium)
```

(d) **(3 pt)** Finally, fill in the blanks below to use `boot_diffs` to compute an approximate 95% confidence interval for the population mean difference between the ratings of Swiss and Belgian chocolate bars. After the code is executed, `left` should store the left endpoint of the interval and `right` should store the right endpoint. You may assume that `boot_diffs` has been computed correctly.

```
left = percentile(2.5, boot_diffs)

right = percentile(97.5, boot_diffs)
```

(e) **(3 pt)** The numerical values of `left` and `right` are -0.03 and 0.56, respectively, so the confidence interval in part (d) is $(-0.03, 0.56)$. Select **ALL** answers that we can justifiably conclude. If you do not have enough information to evaluate whether the statement is true or false, **DO NOT** select it.

○ $(-0.03, 0.56)$ is an approximate 95% confidence interval for the mean difference in the sample.

● If we take 1,000 new random samples of size 1,600 independently from the same population, and construct an approximate 95% bootstrap confidence interval for the mean difference based on each sample, then about 950 of these intervals will include the mean difference in the population.

○ If we construct 1,000 new approximate 95% bootstrap confidence intervals using 5,000 new independent bootstrap samples from the table `chocolate` each time, then about 950 of these intervals will include the mean difference in the population.

(f) **(2 pt)** Suppose we test whether or not Belgian and Swiss chocolate bars have the same mean ratings in the population, using the confidence interval (-0.03, 0.56) from part (e) and a 5% cutoff for the P-value. Pick **ALL** the correct ways to complete the sentence: The test will conclude that the two mean ratings

○ are different ● could be the same ○ are the same

12. (0 points) **Data art (optional)** Draw a graph or picture describing your experience in Data 8.

13. (0 points) Write your name in the space provided on one side of every page of the exam. You're done!