Data 8 Foundations of Data Science
Spring 2018 Final

INSTRUCTIONS

- You have 3 hours to complete the exam.
- The exam is closed book, closed notes, closed computer, and closed calculator, apart from the official final exam reference guide provided with the exam.
- Write each answer **in the space provided for that answer**.
- You can leave all numerical calculations unsimplified.
- Assume that the code statements `import numpy as np` and `from datascience import *` have been executed.

Question 0 (1 point) Write your name and SID in the space provided on one side of every page of the exam.

Last name	
First name	
Student ID number	
CalCentral email (`_@berkeley.edu`)	
GSI	
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. (7 points) Income Intervals

Researchers studying annual incomes in a city take a random sample of 400 households and create 10,000 bootstrap samples from the original sample. They then use the bootstrap percentile method to construct an approximate 95% confidence interval for the median household income in the city. This is the method we have always used in Data 8, and you can assume that it works fine in this situation. The 95% confidence interval goes from $60,000 to $62,000.

(a) (4 pt) Fill in the bubble next to each statement that must be true based on the information above. **More than one might be true**. No explanations are needed.

- ○ About 50% of the households in the city have incomes between $60,000 and $62,000.
- ○ About 95% of the households in the city have incomes between $60,000 and $62,000.
- ○ The researchers are estimating that the median household income in the city is between $60,000 and $62,000, but they could be wrong.
- ○ If the researchers had constructed an approximate 90% confidence interval based on the same bootstrap samples they used for the 95% interval, then both ends of their 90% confidence interval would have been inside the range $60,000 to $62,000.
- ○ None of the above is true.

(b) (3 pt) The array `resampled_medians` contains the 10,000 bootstrapped medians. Complete the code below so that the last line evaluates to the left and right endpoints of an approximate 90% confidence interval for the median household income in the city.

```
left_end = ______________________________________________________________

right_end = _____________________________________________________________

make_array(left_end, right_end)
```

2. (5 points) Sample Size

The table below shows some percentages under the normal curve, in addition to those already in the exam reference guide.

Percent in Range	Normal Distribution
average ± 1.3 SDs	about 80%
average ± 1.65 SDs	about 90%

A researcher wants to estimate the percent of undecided voters in a population, by constructing a confidence interval based on a random sample of voters. Approximately what is the smallest sample size that will ensure that the width of a 90% confidence interval will be no more than 5%?

3. (10 points) Assessing BOKS

In February of this year, the American Journal of Preventive Medicine presented an analysis of a program called BOKS (Build Our Kids' Success) in Massachusetts. BOKS is a before-school exercise program in which children arrive at school one hour early to engage in games and other physical exercise before the day's classes begin.

Participation in BOKS is voluntary. Researchers studied several hundred BOKS participants in kindergarten through eighth grade, and compared their outcomes with those of children who did not join the BOKS program.

Based on measurements taken after 12 weeks of the BOKS program, a smaller proportion of BOKS participants qualified as obese compared to non-participants. Also, BOKS participants reported feeling deeper social connections to their friends and school and greater happiness and satisfaction with life than they did before the program. Non-participants had no changes to their feelings of well being.

The New York Times reported, "The upshot is that a one-hour, before-school exercise program does seem likely to improve young people's health and happiness, says Dr. Elsie Taveras, a professor at Harvard and head of general pediatrics at Massachusetts General Hospital ..."

(a) (2 pt) Was this an observational study or a randomized controlled experiment? Explain your answer.

(b) (2 pt) Was there a treatment group? If so, who was in the treatment group?

(c) (2 pt) Was there a control group? If so, who was in the control group?

(d) (4 pt) Do you think that the researchers' analysis clearly established the effectiveness of BOKS? If you do, then explain why. If you don't, then give the main reason why not, and provide an alternative explanation for the researchers' results.

4. (6 points) Mean Absolute Error

The table `incomes` consists of just one column, labeled `'Income'`. The column contains the incomes of a random sample of workers drawn from a large population. I would like to use this sample to predict the income of a new worker drawn from the population.

To measure how good my prediction is, I will use the mean absolute error instead of the more commonly used mean squared error. For example, if the incomes were \$1, \$2, \$3, and \$4, and I used \$2.50 as my prediction, then the mean absolute error would be

$$\frac{1}{4}\big(|1-2.5|+|2-2.5|+|3-2.5|+|4-2.5|\big)$$

Help me find the best predictor of income based on the table `incomes`, as follows.

Complete the code below so that the last line evaluates to an array whose first element is the predictor that minimizes the mean absolute error and whose second element is the value of the minimum mean absolute error. You should start by defining a function `mae` that computes the mean absolute error. You are not required to use every blank line.

```
def mae(_______________________________):

        _________________________________________________________________________

        _________________________________________________________________________

        _________________________________________________________________________

best_predictor = _______________________________________________________________________

smallest_mae = _________________________________________________________________________

make_array(best_predictor, smallest_mae)
```

5. (11 points) Characteristics of Adults

The table `adults` consists of one row for each adult in a population. Each adult has several attributes including age, hours worked per week, employment status, income bracket, and more. The first few rows of `adults` are shown below. The string '<=50K' stands for the income bracket "at most $50,000".

Age	Education	Relationship Status	Job Sector	Hours per Week	Country	Income
39	Bachelors	Never-married	Adm-clerical	40	United-States	<=50K
50	Bachelors	Married-civ-spouse	Exec-managerial	13	United-States	<=50K
38	HS-grad	Divorced	Handlers-cleaners	40	United-States	<=50K
53	11th	Married-civ-spouse	Handlers-cleaners	40	United-States	<=50K
28	Bachelors	Married-civ-spouse	Prof-specialty	40	Cuba	<=50K
37	Masters	Married-civ-spouse	Exec-managerial	40	United-States	<=50K
49	9th	Married-spouse-absent	Other-service	16	Jamaica	<=50K

Assume that the entries in columns 'Age' and 'Hours per Week' are of type int while the rest are of type string.

In each part below, fill in the blanks of the Python expression to compute the described value. **You must use ONLY the line provided.** The code in the line should evaluate to the value described.

(a) (2 pt) The proportion of adults who work exactly 40 hours a week

```
adults.__________________________________________________ / ________________________
```

(b) (2 pt) The Relationship Status (string) of the youngest person in the table (you can assume there is only one youngest person)

```
adults.____________________________________________________________________________
```

(c) (3 pt) An array consisting of the Education of the oldest adult (or adults, if more than one are oldest)

```
adults.____________________________________________________________________________
```

(d) (4 pt) A two-column table that has one row for each distinct country in the 'Country' column of the table `adults`, such that the first column 'Country' contains the name of the country, and the second column 'Bachelors' contains the average hours per week worked by adults in that country who have a Bachelors degree (and no higher education). For countries where no one has a Bachelors degree, the value of 'Bachelors' should be 0.

```
___________________________________________________________________________________
```

6. (7 points) Estimating a Percentile

Fill in the blanks below with code to define a function `ci_75` that constructs a confidence interval for the 75th percentile of a numerical population, as follows. The function takes the following arguments.

- `tbl`: A one-column table consisting of a random sample from the population; you can assume that the sample is large
- `reps`: a number of bootstrap repetitions; you can assume that users will enter a large integer

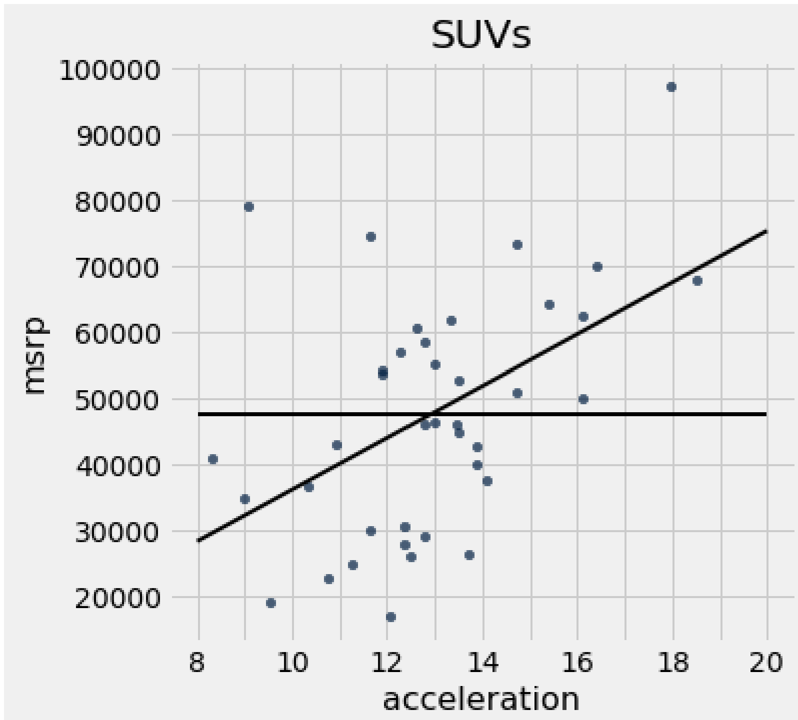
The function returns an array containing the endpoints of an approximate 95% bootstrap confidence interval for the 75th percentile of the population.

```
def ci_75(tbl, reps):

    percentiles = ____________________________________________________

    for _________________________________________________________________

        new_samp = ___________________________________________________________

        new_percentile = _____________________________________________________

        percentiles = ______________________________________________________

    left_end = _____________________________________________________

    right_end = _____________________________________________________

    return make_array(left_end, right_end)
```

7. (13 points) Prediction and Error

In the scatter diagram below, each point represents a model of hybrid SUV. The variables are:

- acceleration, measured in kilometers per hour per second (it doesn't matter if you don't understand those units)
- msrp, an acronym for manufacturer's suggested retail price, in dollars



Some summary statistics:

- The average msrp is \$47,600 and the SD is \$18,000.
- The correlation between msrp and acceleration is 0.5.

The two straight lines:

- The flat line is at the level y = average msrp.
- The slanted line is the regression line for predicting msrp based on acceleration.

(a) **(2 pt)** Fill in the blank and **explain**: If a hybrid SUV is one SD above average in acceleration, the regression prediction of its msrp is \$________________ above the average msrp.
Explanation:

(b) (3 pt) Pick one option and **explain**.
The average acceleration is closest to

- ⬭ 12
- ⬭ 12.5
- ⬭ 13
- ⬭ 13.5
- ⬭ 14

Explanation:

(c) (2 pt) Pick one option to fill in the blank, and **explain**.
To predict msrp based on acceleration, Researcher Ave uses the flat line and Researcher Reg uses the regression line. The root mean squared error made by Researcher Ave is ________________ the root mean squared error made by Researcher Reg.

- ⬭ less than
- ⬭ equal to
- ⬭ greater than

Explanation:

(d) (3 pt) Find the root mean squared error made by Researcher Reg, if it is possible to find it with the information given. **Explain your answer.** If it is not possible to find it, write NA and **explain your choice**.

(e) (3 pt) Find the root mean squared error made by Researcher Ave, if it is possible to find it with the information given. **Explain your answer.** If it is not possible to find it, write NA and **explain your choice**.

8. (6 points) Tests and Error

Researchers are conducting a test of hypotheses using 3% as the cutoff for the P-value.

(a) **(3 pt)** Pick the right option from among the following and **explain your choice**.
If the null hypothesis is true, the chance that the test reaches the correct conclusion is

- ⬭ 100%
- ⬭ 0%
- ⬭ 2%
- ⬭ 3%
- ⬭ 97%
- ⬭ 50%

Explanation:

(b) **(3 pt)** Pick the right option from among the following and **explain your choice**.
The P-value of the test comes out to be 2%. The conclusion of the test is that

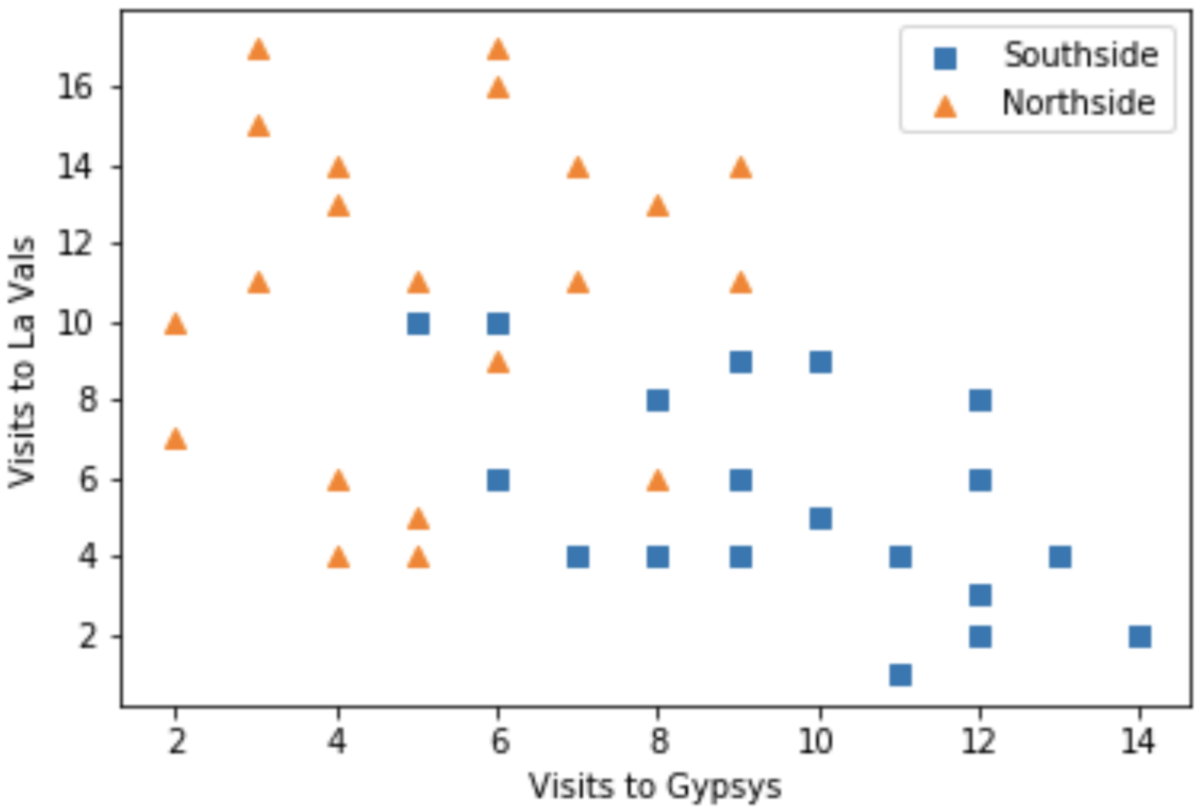
- ⬭ the data support the null hypothesis more than they support the alternative, because there is a 98% chance that the null hypothesis is true.
- ⬭ the data support the alternative hypothesis more than they support the null, because if the null were true then something unlikely has occurred.
- ⬭ the data support the alternative more than they support the null, because there is only a 2% chance that the null hypothesis is true.

Explanation:

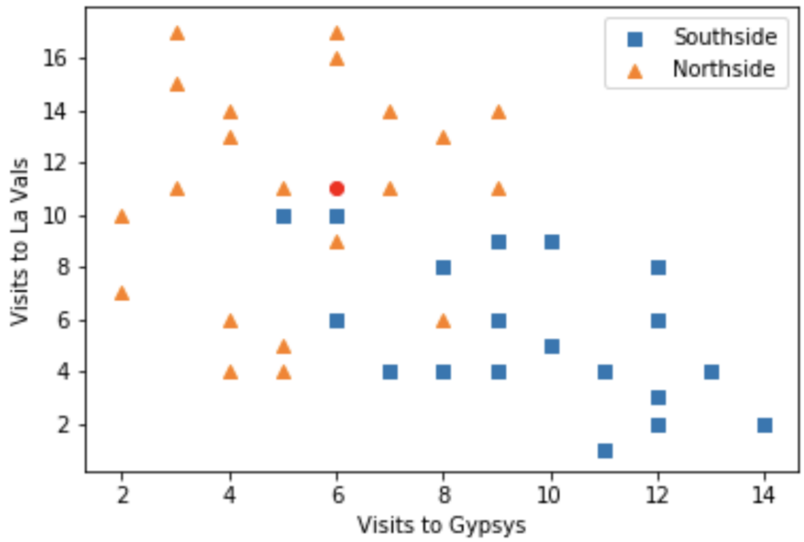
9. (7 points) Northside or Southside

A student is trying to build a classifier that classifies Berkeley students as residents of Northside or Southside. The student has a random sample of Berkeley students all of whom live on Northside or Southside. For each student she records whether the student lives on Northside or Southside, the number of times the student went to La Val's (on Northside) in the last 6 months, and the number of times the student went to Gypsy's (on Southside) in the last 6 months.

(a) (3 pt) Draw an approximate 5-nearest-neighbors decision boundary on the scatter plot of the sample, shown below. No explanation is necessary.



The student is attempting to classify the point at (6, 11), represented by the circle in the graph below.



(b) **(2 pt)** Suppose she uses a 3-nearest-neighbors classifier. What will her classification be? **Explain**.

◯ Northside
◯ Southside

Explanation:

(c) **(2 pt)** Suppose she uses a 5-nearest-neighbors classifier instead. What will her classification be? **Explain**.

◯ Northside
◯ Southside

Explanation:

10. (7 points) Weighted K Nearest Neighbors

Instead of classifying a movie as 'action' or 'romance' according to the class of the majority of its k nearest neighbors in a training set, consider weighting each of the k nearest neighbors by its distance from the unclassified movie. The following method is proposed.

- Find the k nearest neighbors of the unclassified point.
- Let `Largest` be the largest of the distances of these k points from the unclassified point, and let `Smallest` be the smallest of the k distances.
- For each of the k nearest neighbors, define its weight as follows. Let `d` be its distance from the unclassified point. Calculate its weight as
  `(Largest - d)/(Largest - Smallest)`
- Among the k nearest neighbors, find the total weight of the 'action' points as well as the total weight of the 'romance' points.
- Choose the class that has the larger total weight. If the two total weights are equal, choose whichever class you wish.

In order to classify an unclassified movie using the scheme above, complete the definition of a function `weighted_knn` that takes the following arguments:

- `tbl`: A two-column table in which the first column is labeled 'Distance' and the second column is labeled 'Genre'. Each row of the table represents a movie in the training set. The first entry in the row is the distance of the movie from the unclassified movie. The second entry in the row is the genre (either 'action' or 'romance') of the movie.
- `k`: The value of k to use for k-nearest-neighbors

The function returns either 'action' or 'romance' using the weighted k-nearest-neighbors scheme proposed above.

```
def weighted_knn(tbl, k):

    nn = ____________________________________________________

    smallest = min(nn.column('Distances'))

    largest = max(nn.column('Distances'))

    weights_array = ___________________________________________________________________

    weights_tbl = nn.select('Genre').__________________________________________

    by_genre = __________________________________________________________________

    return by_genre.sort(__________________________).column(__________________).item(____)
```

11. (20 points) Voter Distributions

Before a Presidential election, Candidate A's campaign staff studied distributions of voter preferences by taking a random sample of voters in each of two states. Each sampled voter checked one of three boxes:

- Will vote for Candidate A
- Will not vote for Candidate A
- Undecided

The table `voters` contains one row for each sampled voter. The states are labeled State 1 and State 2. Here are the first four rows of `voters`, along with the output of `voters.group('State')`.

State	Preference
State 1	Will not vote for A
State 1	Will vote for A
State 1	Will vote for A
State 1	Undecided

State	count
State 1	2000
State 2	3000

(a) (3 pt) The table `dists`, shown below, contains counts of sampled voters in different categories. For example, there are 540 Undecided voters in the sample from State 1.

Preference	State 1	State 2
Undecided	540	600
Will not vote for A	660	1350
Will vote for A	800	1050

Fill in the blank so that the line of code produces the table `dists`.

```
dists = voters.__________________________________________________________________________
```

(b) (4 pt) Are the distributions of preference different in the two states? State the null and alternative hypotheses that should be used to answer this question.

Null:

Alternative:

(c) **(1 pt)** What test statistic should be used to test the hypotheses in (b)?

(d) **(1 pt)** Complete the line of code below so that `obs_stat` is the observed value of the test statistic. You can use the table `dists` in your code.

```
obs_stat = ___________________________________________________________________
```

(e) **(3 pt)** Explain how to simulate the test statistic under the null hypothesis. You will write the code in the next part.

(f) **(4 pt)** Complete the definition of the function `simulate_stat` so that it returns one value of the test statistic simulated under the null hypothesis. The function takes no argument.

```
def simulate_stat():

       new_array = voters.__________________________________.column(____________________)

       new_tbl = voters.select('State').____________________________________________

       dists = new_tbl._______________________________________________________________

       return ________________________________________________________________________
```

(g) **(4 pt)** Fill in the code so that the last line evaluates to the P-value of the test based on 10,000 simulated values of the test statistic. Use `simulate_stat` and `obs_stat` in your code.

```
stats = ______________________________________________________________________

for ________________________________________________

       stats = _________________________________________________________________

__________________________________________________________________________________________
```