Data 100 & 200A Principles and Techniques of Data Science
Spring 2019

MIDTERM 1 SOLUTIONS

INSTRUCTIONS

- You have 70 minutes to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except one hand-written 8.5" × 11" crib sheet of your own creation and the official Data 100 study guide.
- Mark your answers **on the exam itself**. We will *not* grade answers written on scratch paper.

Last name	
First name	
Student ID number	
CalCentral email (`_@berkeley.edu`)	
Exam 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)**	

1. (12 points) Python vs. SQL

Fill in both the Python code and the SQL query to produce each result below, assuming that the following three tables are stored both as Pandas DataFrames and Sqlite tables. **Only the first few rows are shown for each table.** The `cities` table contains one row per city and its population in thousands. The `names` table contains one row per state. The `states` table contains one row per state with its population in millions. Assume that `cities` contains only a small subset of US cities. There may be multiple cities in each state, but every city is in a state that appears in both `names` and `states`, and every state contains at least one city.

`cities`

city	pop
Nevada City, California	3
Carson City, Nevada	55
Newark, New Jersey	285

`names`

abbrev	full
CA	California
NV	Nevada
WA	Washington

`states`

state	people
CA	39.5
NV	3.0
WA	7.4

(a) (4 pt) Create a table `t` that is the same as `cities` but with an additional column `ab` containing the abbreviation (e.g. CA) of the state in which each city is located. *Hint*: The `str.extract` method of a Series called on a regular expression with one group returns a DataFrame with one column labeled 0 containing the first substring matching the group. Assume there is exactly one comma in each `city` value.

Python:
```
t = cities.copy()

x = cities['city'].str.extract(r', (\w+)')[0]

t['ab'] = list(names.set_index('full').loc[x, 'abbrev'])
```

SQL:
```
CREATE TABLE t AS SELECT city, pop, abbrev AS ab FROM cities JOIN names ON

city LIKE '%, ' || full;
```

(b) (4 pt) Create a two-column table `u` of the cities and their populations (labeled `city` and `pop`) that are in states with a population above 5 million. Assume that `t` from part (a) was constructed correctly.

Python:
```
m = t.merge(states, left_on='ab', right_on='state')

m[m['people'] > 5][['city', 'pop']]
```

SQL:
```
CREATE TABLE u AS SELECT city, pop FROM t WHERE

ab IN (SELECT state FROM states WHERE people > 5)
```

(c) (4 pt) Create a table with one row per state that contains the state's abbreviation and the fraction of cities (from the `cities` table) in that state that have a population above 50,000.

Python:
```
t['pop'].groupby(t['ab']).agg(

 lambda s: sum(s>50)/len(s))
```

SQL:
```
SELECT ab, SUM(CASE WHEN pop < 50 THEN 0 ELSE 1 END)/COUNT(*)

      FROM t GROUP BY ab;
```

2. (6 points) Sampling

Circle the correct response to each question about this population of six individuals.

name	cluster
Abdul	Blue
Ace	Blue
Adele	Blue
Aerie	Blue
Bella	Gold
Buzz	Gold

(a) (2 pt) From the population above, you draw a simple random sample A of 2 individuals. What's the probability that Ace and Adele both appear in sample A?

$\frac{1}{6} \cdot \frac{1}{6}$ $\frac{1}{6} \cdot \frac{1}{5}$ $\frac{1}{3} \cdot \frac{1}{3}$ $\frac{1}{6} + \frac{1}{6}$ $2 \cdot \frac{1}{6} \cdot \frac{1}{6}$ $2 \cdot \frac{1}{6} \cdot \frac{1}{5}$ $2 \cdot (\frac{1}{3} \cdot \frac{1}{3})$ $2 \cdot (\frac{1}{6} + \frac{1}{6})$

(b) (2 pt) You separately draw a cluster sample B from the same population based on the `cluster` column. (Sample A is replaced in the population before drawing sample B, so the two are independent.) What's the probability that Ace and Adele both appear in sample B?

0 $\frac{1}{6}$ $\frac{1}{4}$ $\frac{1}{3}$ $\frac{1}{2}$ $\frac{2}{3}$ $\frac{3}{4}$ None of these

(c) (2 pt) You then combine all individuals from sample A and sample B into sample C. Thus, there may be repeated individuals in sample C. What's the probability that Bella appears exactly once in sample C?

0 $\frac{1}{6}$ $\frac{1}{4}$ $\frac{1}{3}$ $\frac{1}{2}$ $\frac{2}{3}$ $\frac{5}{6}$ None of these

3. (4 points) Regular Expressions

```
[a-z]+_[a-z]{2}[_r]_?[a-z]+
```

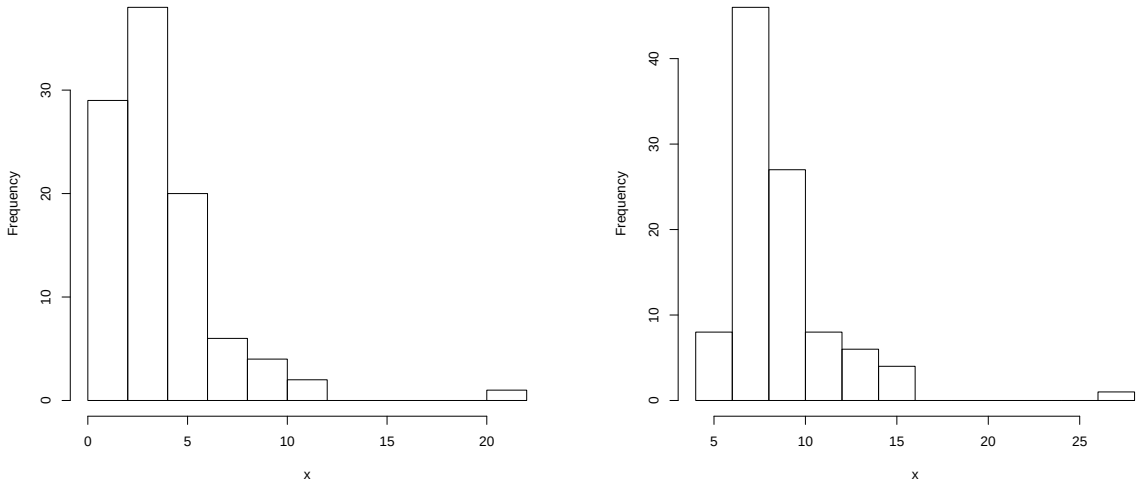
Circle **all** of the strings below that match the regular expression above. Only circle a string below if the **whole string** matches the expression, not just a substring.

`bar_chart` `group_by_x` `a_box_plot` `visualize_first`

4. (10 points) Data Visualization

(a) (2 pt) Are the two histograms below displaying exactly the same data? Circle only one answer.

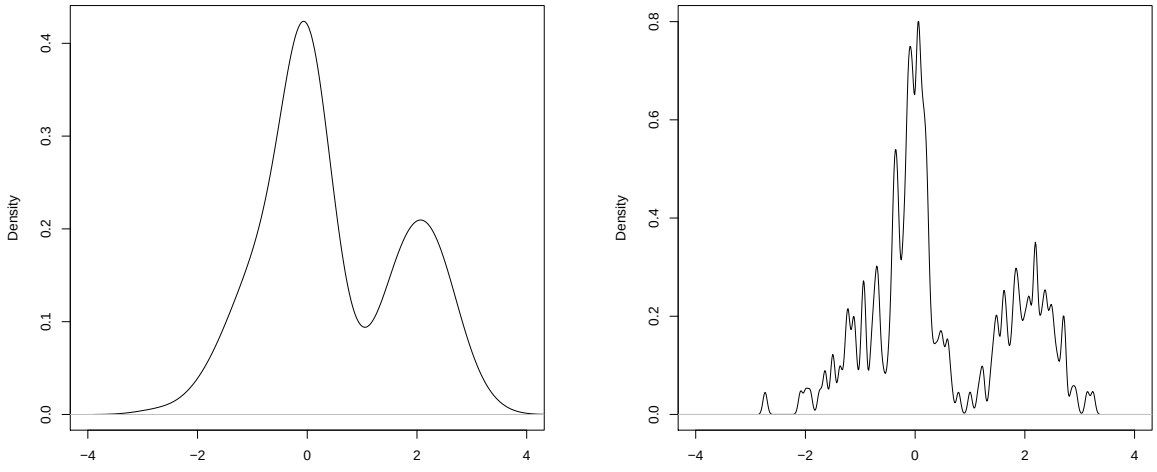
(a) Yes (b) No (c) Impossible to tell



There is a location shift between the two distributions.

(b) (2 pt) Are the two Gaussian kernel density plots below displaying exactly the same data? Circle only one answer.

(a) Yes (b) No (c) Impossible to tell

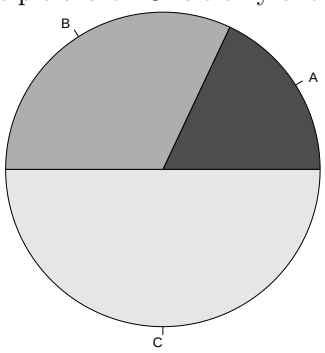


The density plots could be displaying the same data and look very different because of the bandwith, however, one cannot tell for sure.

(c) (2 pt) Which of the following can be determined from looking at this pie chart? Circle only one answer.

(a) Category B is twice as frequent as category A
(b) Category A is half as frequent as category B
(c) Category B is more frequent than category A
(d) All of the above
(e) None of the above

With angles/areas, it is very hard to precisely compare frequencies.



(d) (2 pt) You have 30 unique observations of a real-valued quantitative variable. Which of the following visualizations effectively depicts the distribution of these values while retaining as much information as possible about the original data? Circle only one answer.

(a) A pie chart
(b) A strip chart
(c) A box plot

A strip chart displays all of the data. A boxplot would summarize the data and potentially miss issues such as bimodality. A pie chart is not applicable.

(e) (2 pt) You are given six lists, each of a few thousand numbers taking on values in the real line. Which of the following is the most effective way to visually compare the center and spread of the corresponding six distributions? Choose only one answer.

(a) Side-by-side box plots
(b) Side-by-side histograms
(c) Side-by-side strip charts

With a few thousand values, strip charts would be unreadable. Visually comparing several histograms is difficult.

5. (8 points) Dimensionality Reduction

You perform principal component analysis on a data matrix D using the following Python code from lecture:

```
n = D.shape[0]
X = (D - np.mean(D, axis=0)) / np.sqrt(n)
u, s, vt = np.linalg.svd(X, full_matrices=False)
```

The resulting value of `s` is `np.array([3, 1, 0, 0, 0])`.

(a) (4 pt) To draw a histogram of the data's distribution along the first principal component of X, which of the following arrays would you visualize? Circle **all** correct expressions.

`X @ u.T[:,0]` `(u * s)[:,0]` `X @ vt[0,:]` `(X @ vt.T)[:,0]`

(b) (2 pt) What proportion of the total variance in D is accounted for by the first principal component?

$\frac{9}{10}$

(c) (2 pt) What is the rank of X?

2