DS-100 Practice Midterm Questions

Spring 2017

Note: The following questions are intended to be representative of what you'll see on the midterm. The actual exam will have a single page (front and back) answer sheet on which to write each of the answers. The following questions are not guaranteed to cover every topic that's fair game for the exam, and this set is not representative of the length of the exam.

1. True or False

   (1) [1 Pt.] All data science investigations start with an existing dataset.

   (2) [1 Pt.] Because smoking is viewed as a cause for lung cancer, it does not make sense to use lung cancer status to predict smoking status.

   (3) It is possible that the null hypothesis will be rejected when it is in fact true.

   (4) It is possible that the alternative hypothesis is true when the null hypothesis is not rejected.

   (5) Practically significant effects will always yield statistically significant results.

2. Consider two relations with the same schema: `R(A,B)` and `S(A,B)`. Which one of the following relational algebra expressions is not equivalent to the others?

   A. $\pi_{R,A}((R \cup S) - S)$

   B. $\pi_{R,A}R - \pi_{R,A}(R \cap S)$

   C. $\pi_{R,A}(R - S) \cap \pi_{R,A}R$

   D. They are all equivalent.

3. Consider the following real estate schema:

```
Homes(home_id int, city text, bedrooms int, bathrooms int,
      area int)
Transactions(home_id int, buyer_id int, seller_id int,
             transaction_date date, sale_price int)
Buyers(buyer_id int, name text)
Sellers(seller_id int, name text)
```

For the query language questions below, fill in the blanks in the answer to complete the query. For each SQL query and nested subquery, please start a new line when you reach a SQL keyword (SELECT, WHERE, AND, etc.). However, do not start a new line for aggregate functions (COUNT, SUM, etc.), and comparisons (LIKE, AS, IN, NOT IN, EXISTS, NOT EXISTS, ANY, or ALL.)

(1) Fill in the blanks in the SQL query to find the duplicate-free set of id's of all homes in Berkeley with at least 6 bedrooms and at least 2 bathrooms that were bought by "Bobby Tables."

SELECT ____

FROM ____

WHERE ____

____

____

____

____

____

(2) Repeat the above query using relational algebra:

$\pi\underline{\hspace{3cm}}($

$\quad\sigma\underline{\hspace{7cm}}($

$\quad((\sigma\underline{\hspace{14cm}}\underline{\hspace{2.5cm}})$

$\quad\;\bowtie\underline{\hspace{3.5cm}})\bowtie\underline{\hspace{2.5cm}}))$

(3) Fill in the blanks in the SQL query to find the id and selling price for each home in Berkeley. If the home has not ben sold yet, **the price should be NULL**.

```
SELECT ______________________________
FROM __________________________________
_________________ JOIN _________________
ON _________________________
WHERE __________________________;
```

4. A biologist wants to figure out how a set of organisms are related to each other, so she runs an algorithm that divides the species into 3 groups based on their traits. This is an example of (choose all that apply):

   A. supervised learning

   B. unsupervised learning

   C. clustering

   D. classification

5. Chinua decides to investigate the effect of an IQ-enhancing drug, Versivium, on Berkeley students. He magically procures a list of all the students at Berkeley and their contact information. After randomizing the order of the list, he selects the top 1000 to invite to his experiment. Miraculously, all invitees agreed to be part of this clinical trial—anything for that A+. He divides the students into two groups: 800 in control and 200 in treatment. Members of the treatment group were administered a dose of Versivium every day for 20 days. Their counterparts in the control group were given an identical-looking sugar pill in the same dosage. All participants were compliant. At the end of the trial, the participants were asked to take an IQ test. When Chinua performed a test against the null hypothesis that the two groups had the same mean IQ score, he found that the difference in test scores was significant at the 5% cutoff. Which of the following are valid statements?

   A. The results of the test are actually inconclusive since the the control and treatment groups were uneven.

   B. We should be doubtful of the results since Chinua took the top 1000 names of his ordered list.

   C. The experiment is not well-designed as 1000 is too small a sample to make conclusions about the whole student body

D. It could be the case that we were unlucky and the difference in IQ scores observed were due to chance

E. None of the above

6. [3 Pts.] Consider the following plot about how baby boomers describe themselves. Which mistakes does it make? Circle all that apply.

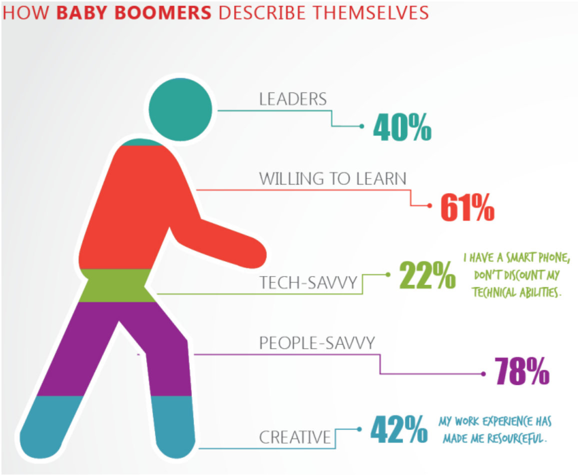
A. sampling bias

B. jiggling base line

C. stacking

D. jittering

E. area perception



7. [3 Pts.] The FEC data includes contributions to the Clinton and Sanders campaigns. If we want to create a visualization that helps us compare the sizes of donations to their campaigns, which of the following plots should we make? Circle all that apply.

A. scatter plot with donations to Clinton's campaign on one axis and Sanders' on the other.

B. density curve of Clinton donations over laid on density curve of Sanders donations.

C. side-by-side bar plot of their donations

D. Two box plots, one for Clinton donations and one for Sanders.

E. None of the above

8. [3 Pts.] Maximize the following likelihood with respect to p

$$L(p) = (k_1 k_2 k_3) p^6 (1-p)^{(k_1+k_2+k_3)-6}$$

Note, $\bar{k} = (k_1 + k_2 + k_3)/3$.

A. $\bar{k}/2$

B. $2/\bar{k}$

C. $-2/\bar{k}$

D. $2/(4+\bar{k})$

E. $2/(4-\bar{k})$

9. Suppose X, Y, and Z are random variables that are independent and have the same probability distribution. If $\mathrm{Var}(X) = \sigma^2$, then $\mathrm{Var}(X+Y+Z)$ is:

   A. $9\sigma^2$

   B. $3\sigma^2$

   C. σ^2

   D. $\frac{1}{3}\sigma^2$

10. A jar contains 3 red, 2 white, and 1 green marble. Aside from color, the marbles are indistinguishable. Two marbles are drawn at random without replacement from the jar. Let X represent the number of red marbles drawn.

    (1) [2 Pts.] What is $\mathbb{P}(X=0)$?

    A. $1/9$

    B. $1/5$

    C. $1/4$

    D. $2/5$

    E. none of the above

    (2) [2 Pts.] let Y be the number of green marbles drawn. What is $\mathbb{P}(X=0, Y=1)$?

    A. $\frac{1}{15}$

    B. $\frac{2}{15}$

    C. $\frac{1}{12}$

    D. $\frac{1}{6}$

    E. $\frac{7}{15}$

    F. $\frac{8}{15}$

11. [3 Pts.] Suppose the random variable X can take on values -1, 0, and 1 with chance p^2, $2p(1-p)$ and $(1-p)^2$, respectively, for $0 \le p \le 1$. We observe a sequence of 3 independent observations from this distribution. They are $1, 1, 0$.

    What is the likelihood for p?

    A. $2p(1-p)^3$

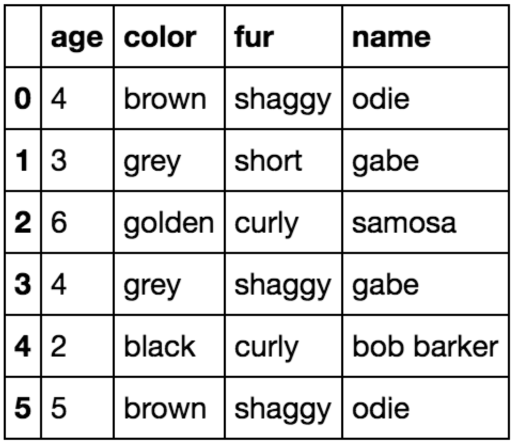
    B. $p^2(1-p)^2$

    C. $(1-p)^4$

    D. $2p(1-p)^5$

    E. $2p^3(1-p)^3$

12. [8 Pts.] The pandas dataframe *dogs* contains information on pets' visits to a veterinarian's office. A portion of the dataframe is shown below.



	age	color	fur	name
0	4	brown	shaggy	odie
1	3	grey	short	gabe
2	6	golden	curly	samosa
3	4	grey	shaggy	gabe
4	2	black	curly	bob barker
5	5	brown	shaggy	odie

For each question, provide a snippet of pandas code as your solution. Assume that the table *dogs* has the same column format as the provided table (just more rows).

(1) How many different dogs visited the veterinarian's office? Provide code that outputs the answers as an integer. Assume that no two dogs have the same name.

A. `len(dogs.groupby("name").count())`

B. `len(dogs["name"])`

C. `len(dogs)`

(2) What was the name of the oldest dog that visited the veterinarian's office?

A. `dogs.sort_values("age", ascending=False).name[0]`

B. `dogs.sort_values("age", ascending=False).name.iloc[0]`

C. `dogs.groupby("name").agg({"age": "max"})`

(3) What was the most common fur color among dogs?

A.
```
dogs.groupby("color").count().sort_values("name",
    ascending=False).index[0]
```

B.
```
dogs.groupby("color").count().sort_values("age",
    ascending=False).index[0]
```

C.
```
dogs.groupby("color").count().sort_values("fur",
    ascending=False).index[0]
```

D. All of the above.

E. None of the above.

(4) What proportion of dogs had the most common fur type? (For instance, if the most common fur type was curly, what proportion of dogs had curly fur)?

A.
```
dogs.groupby("fur").count().sort_values("age",
    ascending=False).age.iloc[0]/len(dogs)
```

B.
```
dogs.groupby("fur").count().sort_values("age",
    ascending=False).age.iloc[0]/len(dogs["age"])
```

C. `dogs.groupby("fur").count().sort_values("age", ascending=False).age.iloc[0]/len(dogs["fur"])`

D. All of the above.

E. None of the above.