# DS-100 Practice Midterm Questions

Spring 2018

**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.

# Contents

# Syntax Reference

## Regular Expressions

**`"^"`** matches the position at the beginning of string (unless used for negation `"[^]"`)

**`"$"`** matches the position at the end of string character.

**`"?"`** match preceding literal or sub-expression 0 or 1 times. When following `"+"` or `"*"` results in non-greedy matching.

**`"+"`** match preceding literal or sub-expression *one* or more times.

**`"*"`** match preceding literal or sub-expression *zero* or more times

**`"."`** match any character except new line.

**`"[ ]"`** match any one of the characters inside, accepts a range, e.g., `"[a-c]"`.

**`"( )"`** used to create a sub-expression

**`"\d"`** match any *digit* character. `"\D"` is the complement.

**`"\w"`** match any *word* character (letters, digits, underscore). `"\W"` is the complement.

**`"\s"`** match any *whitespace* character including tabs and newlines. `\S` is the complement.

**`"\b"`** match boundary between words

Some useful `re` package functions.

**`re.split(pattern, string)`** split the `string` at substrings that match the `pattern`. Returns a list.

**`re.sub(pattern, replace, string)`** apply the `pattern` to `string` replacing matching substrings with `replace`. Returns a string.

## Useful Pandas Syntax

```
df.loc[row_selection, col_list]  # row selection can be boolean
df.iloc[row_selection, col_list] # row selection can be boolean
df.groupby(group_columns)[[’colA’, ’colB’]].sum()
pd.merge(df1, df2, on=’hi’) # Merge df1 and df2 on the ’hi’ column
```

```
pd.pivot_table(df,                    # The input dataframe
               index=out_rows,        # values to use as rows
               columns=out_cols,      # values to use as cols
               values=out_values,     # values to use in table
               aggfunc="mean",        # aggregation function
               fill_value=0.0)        # value used for missing comb.
```

# 1 Data Science Basics

1. True or False

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

**Solution:** **False.** In many settings a data scientist is tasked with a question or problem and must decide how to collect or obtain data to answer the question or solve the problem.

(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.

**Solution:** **False.** Lung cancer is likely a very good predictor of smoking habits. Just because there is no causal relationship (e.g., lung cancer doesn't cause smoking) does not mean that we cannot use one to predict the other.

(3) [1 Pt.] Exploratory data analysis is the process of testing key hypotheses.

**Solution:** **False.** Exploratory data analysis is the process of gaining understanding about data to inform future analysis.

(4) [1 Pt.] In most data science applications only a small amount of time is spent cleaning and preparing data.

**Solution:** **False.** Sadly, studies have shown that a large fraction of time is spent preparing and cleaning data for subsequent analysis.

# 2 SQL

2. 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 DISTINCT H.home_id
FROM Homes H, Transactions T, Buyers B
WHERE H.home_id=T.home_id
AND T.buyer_id=B.buyer_id
AND H.city="Berkeley"
AND H.bedrooms>=6
AND H.bathrooms>=2
AND B.name='Bobby Tables';
```

(2) 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 been sold yet, **the price should be NULL**.

```
SELECT H.home_id, T.sale_price
FROM Homes H
LEFT OUTER JOIN Transactions T
ON H.home_id = T.home_id
WHERE H.city = 'Berkeley';
```

**Solution:** An alternate solution was to use Transactions in the FROM clause and perform a RIGHT OUTER JOIN with Homes.

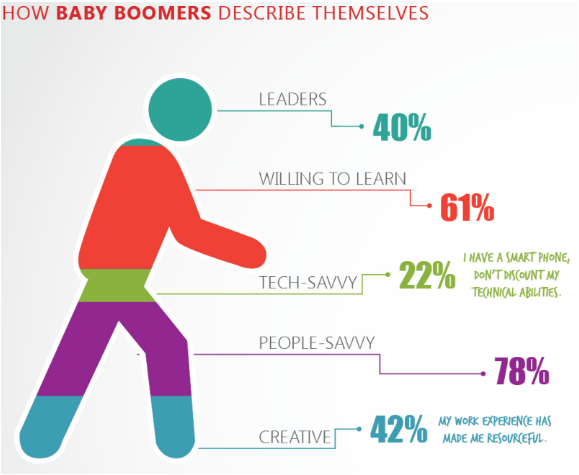
```
SELECT H.home_id, T.sale_price
FROM Transactions T
```

```
RIGHT OUTER JOIN Homes H
ON H.home_id=T.home_id
WHERE H.city = 'Berkeley'
```

# 3 Data Visualization

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

□ sampling bias □ jiggling base line √ **stacking** □ jittering √ **area perception**



4. [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? Select all that apply.

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

   √ **density curve of Clinton donations over laid on density curve of Sanders donations.**

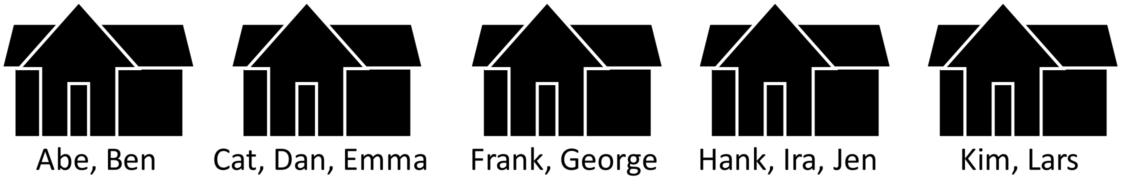
   □ side-by-side bar plot of their donations

   √ **Two box plots, one for Clinton donations and one for Sanders.**

   □ None of the above

# 4 Sampling

5. A small town has 5 houses with the following people living in each house:



Suppose we take a **cluster sample** of 2 houses (without replacement), what is the chance that:

(1) [2 Pts.] Kim and Lars are in the sample

○ 0 ○ 1/20 ○ 1/10 ○ 1/6 ○ 1/5 [✓ **2/5**] ○ 1

**Solution:** The chance that Kim and Lars are in the same sample is given by the chance of choosing their house. The chance of choosing the their house on the first draw is $\frac{1}{5}$. Because we are drawing without replacement. The chance of choosing their house on the second draw is given by the chance of not choosing their house on the first draw ($\frac{4}{5}$) times the chance of choosing their house on the second draw ($\frac{1}{4}$). Thus the total chance of choosing them in the first two draws is:

$$\frac{1}{5} + \frac{4}{5} \times \frac{1}{4} = \frac{2}{5}$$

(2) [2 Pts.] Kim, Abe, and Ben are in the sample

○ 0 ○ 1/20 [✓ **1/10**] ○ 1/6 ○ 1/5 ○ 2/5 ○ 1

**Solution:** To draw Kim, Abe, and Ben we would need to draw both of their houses. This can be done two ways (draw Abe and Ben's house first and then Kim's or vice versa). Each way has probability:

$$\frac{1}{5} \times \frac{1}{4}$$

Thus the total probability is:

$$2 \times \frac{1}{5} \times \frac{1}{4} = \frac{2}{20} = \frac{1}{10}$$

(3) [1 Pt.] Kim and Dan are in the sample - **Select all that apply**

□ The same as the chance Kim and Lars are in the sample

√ **The same as the chance Kim, Abe, and Ben are in the sample**

□ Neither of the above

# 5 Probability

6. 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 $\mathbb{P}(X = 0)$?

○ $1/9$ ✓ $1/5$ ○ $1/4$ ○ $2/5$ ○ none of the above

**Solution:** The event that $X = 0$ is the same as the event that no red marbles are drawn, which is the same as the event that the first draw isn't red and the second draw isn't red.

$p = P(\text{first draw is not red and second draw is not red})$
$= P(\text{first draw is not red})P(\text{second draw is not red — first draw is not red})$

If the first draw isn't red, there are 5 marbles left, 3 of which are red, so:

$$p = \frac{1}{2}\frac{2}{5} = \frac{1}{5}$$

.

A more brute-force counting argument is as follows. There are $\binom{6}{2} = \frac{6!}{4!2!} = 15$ ways to draw a subset of 2 marbles. Of those, the number of subsets with no red marbles is $\binom{3}{2} = \frac{3!}{2!1!} = 3$, so the proportion of draws without red marbles is $3/15 = 1/5$. However it's probably better to exercise your probabilistic thinking via the previous solution!

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

○ $\frac{1}{15}$ ✓ $\frac{2}{15}$ ○ $\frac{1}{12}$ ○ $\frac{1}{6}$ ○ $\frac{7}{15}$ ○ $\frac{8}{15}$

**Solution:** For $X$ to be 0 and $Y$ to be 1, means that we drew 1 green and 1 white ball. We can draw green first and then white, which has chance $1/6 \times 2/5$ or white first and green second, which has chance $2/6 \times 1/5$. The combined probability is $4/30$ or $2/15$.

Another approach is to use conditional probability, i.e.,

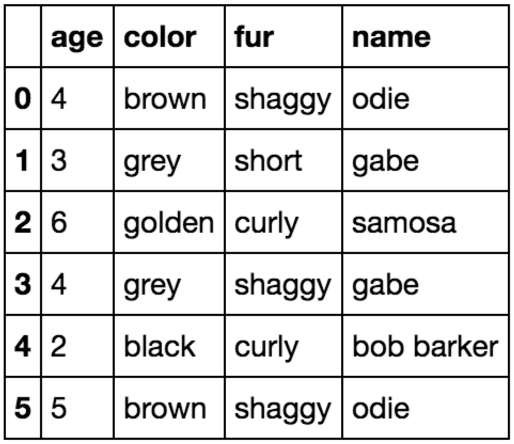
$$\mathbb{P}(X = 0, Y = 1) = \mathbb{P}(X = 0)\mathbb{P}(Y = 1|X = 0).$$

We found $\mathbb{P}(X = 0)$ above to be $1/5$. For the conditional probability, if we know $X = 0$ then we know that we are drawing from the 2 white and 1 green marbles. There are 3 possible ways to draw 2 marbles from these 3 and 2 of the possibilities give us 1 green and 1 white. Putting these together we have $1/5 \times 2/3 = 2/15$.

Alternatively, we can brute-force count the number of subsets that have 1 green and one white marble, which is 2, and divide by the number of ways to choose 2 marbles out of 6 (which we calculated above to be 15).

# 6 Pandas

7. [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.

√ `len(dogs.groupby("name").count())`

○ `len(dogs["name"])`

○ `len(dogs)`

**Solution:** Note that the second and third choices do not account for duplicate appearances by the same name.

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

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

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

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

**Solution:** The first solution would return the dog which had pandas index 0 (that is, the one that appeared in the first row of the dataframe *before* sorting). The third solution returns the maximum age recorded for each dog, but doesn't choose the oldest among them.

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

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

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

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

□ 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)?

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

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

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

□ None of the above.

# 7 Regular Expressions

8. [2 Pts.] Select **all** the strings that **fully match** the regular expression: **`[^dp]an`**

   √ **`Dan`** □ `pan` √ **`fan`** √ **`man`** □ None of the above.

9. [2 Pts.] Select **all** the strings that **fully match** the regular expression: <**`[a-z]*@\w+.edu`**>

   □ `<xin.wang@berkeley.edu>`

   √ **`<@berkeley$edu>`**

   √ **`<xinwang@berkeley#edu>`**

   □ `<xinwang@.edu>`

   □ None of the above strings match.

10. [2 Pts.] Select **all** the strings that **fully match** the regular expression: **`^Go.*`**

    □ `Way to ^Go!`

    √ **`Go Bears!`**

    □ `go trees?`

    □ None of the above strings match

11. [2 Pts.] What is the result of evaluating the following python command?

    ```
    len(re.split(r"\d+", "You get a 99.9 on the exam."))
    ```

    ○ 2 √ **3** ○ 4 ○ 5

12. For the following tasks, write the corresponding Python code or regular expression.

    (1) [2 Pts.] Write a regular expression that only matches sub strings consisting of an `a` immediately followed by zero or one `b` characters.

    `regx = r'`________________________________________`'`

    **Solution:**

    ```
    regx = r'ab?'
    ```

    (2) [3 Pts.] Suppose we've run the code below:

    ```
    text = 'Data\t \t Science  100'
    ```

Use a method in the `re` module to replace all the continuous segments of spaces with a single comma. The resulting string should look like "`Data,Science,100`".

`re.`________________________________________________________

**Solution:**

```
re.sub(r'\s+', ',', text)
```

# 8 Modeling and Loss Minimization

13. [6 Pts.] We propose the following simple model for a dataset consisting of four points $\mathcal{D} = \{0, 2, 4, 10\}$:

$$y = \theta^*$$

Use the following plots of loss functions for this model to answer the following questions.

The plot on the left shows the average squared loss versus $\theta$; the plot on the right shows the average absolute loss.

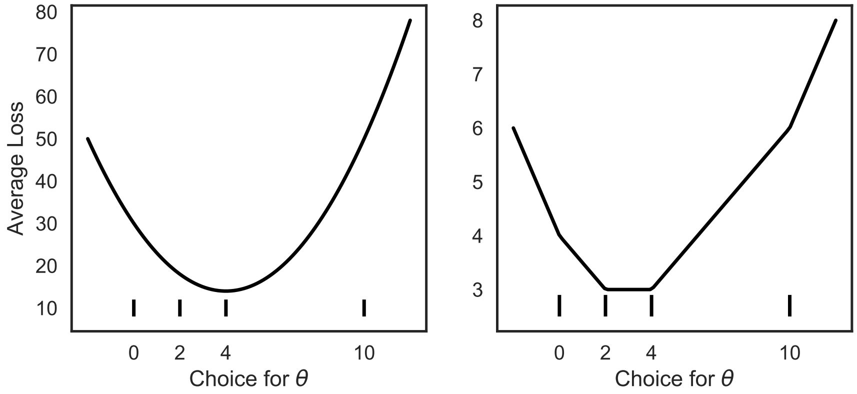


Figure 1:

(1) [2 Pts.] Which choice(s) for $\theta$ minimize the average squared loss? **Select all that apply.**

□ 2 □ 3 √ **4** □ 10 □ None of the above

(2) [2 Pts.] Which choice(s) for $\theta$ minimize the average absolute loss? **Select all that apply.**

√ **2** √ **3** √ **4** □ 10 □ None of the above

(3) [2 Pts.] Suppose we add an observation at $y_5 = 100$. Which choice(s) for $\theta$ minimize the average absolute loss? **Select all that apply.**

A. A value smaller than 3
B. 3
**C. 4**
D. 5
E. A value larger than 5

14. Which $\theta$ minimizes the following loss function for a dataset $D$ comprised of $(x_i, y_i)$ pairs? **Show your work in the space provided.**

$$L(\theta, D) = \sum_{i=1}^{n} (y_i - \theta x_i)^2$$

○ $\theta = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{x_i}$ ✓ $\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ ○ $\theta = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$ ○ $\theta = \frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i^2}$

15. Suppose we observe a dataset $\{x_1, \ldots, x_n\}$ and the following loss function for the parameter $\lambda$:

$$L(\lambda, \mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \ln(\lambda e^{-\lambda x_i})$$

Derive the loss minimizing parameter value $\hat{\lambda}$. **Circle your answer.**

**Solution:** Taking the derivative of the loss function with respect to the parameter $\lambda$ we get:

$$\begin{aligned}
\frac{\partial}{\partial \lambda} L(\lambda, \mathcal{D}) &= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \lambda} \ln\left(\lambda e^{-\lambda x_i}\right) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \lambda} \left(\ln(\lambda) + \ln\left(e^{-\lambda x_i}\right)\right) && (1)\\
&= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \lambda} \left(\ln(\lambda) - \lambda x_i\right) = -\frac{1}{n}\sum_{i=1}^{n} \left(\frac{1}{\lambda} - x_i\right) && (2)\\
&= -\frac{1}{\lambda} + \frac{1}{n}\sum_{i=1}^{n} x_i && (3)
\end{aligned}$$

To compute the loss minimizing parameter $\hat{\lambda}$ we set the above derivative equal to zero and solve.

$$\begin{aligned}
0 &= -\frac{1}{\lambda} + \frac{1}{n}\sum_{i=1}^{n} x_i && (4)\\
\frac{1}{\lambda} &= \frac{1}{n}\sum_{i=1}^{n} x_i && (5)\\
\lambda &= \frac{n}{\sum_{i=1}^{n} x_i} && (6)
\end{aligned}$$

Thus the loss minimizing parameter estimate is:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1} = \frac{1}{\textbf{Mean}(x)} \qquad (7)$$

# 9 Web Technologies

16. [1 Pt.] HTTP is a simple _____ protocol.

    A. Push - Pull

    B. Get - Post

    **C. Request - Response**

    D. Read - Write

17. [3 Pts.] Data100 grade-book server can do two things:

    1. Reader can add/update student's grade on the server.
    2. Students can retrieve their grade but cannot modify them.

    What's the appropriate REST request type for adding/updating grade?

    A. GET

    **B. POST**

    C. OPTIONS

    What's the appropriate REST request type for retrieving grade?

    **A. GET**

    B. POST

    C. OPTIONS

    Reader can submit multiple grades in JSON format. Given the following JSON, select all the true statements:

```
[{
    "student_id": 1,
    "assignment_id": 23,
    "grades": {
        "q1": 2,
        "q2": 3
    },
    "comments": "great plot!",
    "comments": "nice explanation!"
}, {
    "student_id": 1,
    "assignment_id": 24,
    "grades": {
        "q1": "20",
16         "q2": "10"
17     }
18 }]
```

A. The square bracket at line 1 and 18 are redundant. They should be taken out.

**B. Duplicate keys on line 8 and 9 are not allowed.**

C. All the integer should be string type.

D. “grades” field should be a list of grade not a nested JSON.