DS-100 Practice Midterm Exam Questions

Fall 2017

Name: ____

Email address: ______________________________________

Student id: ____

Instructions:
This is a collection of practice questions for the midterm exam.

Syntax Reference

On the exam we will provide **this** reference sheet for basic syntax.

Regular Expressions

"^" matches the position as 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

```
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.

df.groupby(group_columns)[['colA', 'colB']].sum()
df.loc[row_selection, col_list] # row selection can be boolean
```

1. True or False

(1) 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) Data scientists do most of their work in Python and are unlikely to use other tools.

> **Solution:** **False.** Data scientists use many programming languages and tools. In class we discussed surveys that suggested that SQL and then R are the most commonly used languages.

(3) Most data scientists spend the majority of their time developing new models.

> **Solution:** **False.** Sadly, data suggests that most data scientists spend the majority of their time collecting and cleaning data and doing exploratory data analysis.

(4) The use of historical data to make decisions about the future can reinforce historical biases.

> **Solution:** **True.** A key ethical challenge of data driven decision making is that we tend to reinforce trends in our data.

(5) Using properly constructed statistical tests, it is possible that the null hypothesis will be rejected when it is in fact true.

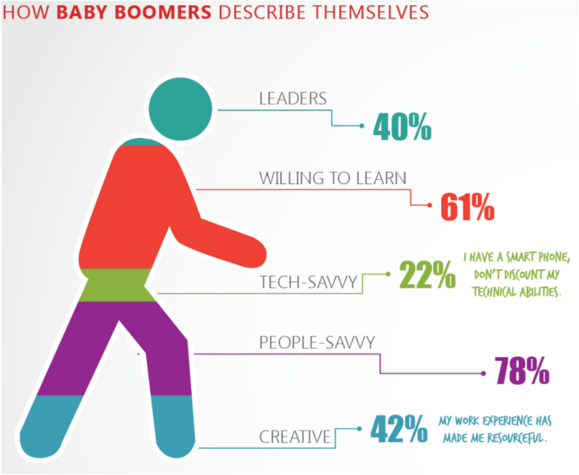
> **Solution:** **True.** We reject the null hypothesis when the chance of observing data/statistics like ours is very small, but this means that we may be erroneously rejecting the null hypothesis. That is, we may have observed a rare event under the null model, and we are rejecting it even though it is true.

(6) Bootstrapping ‘works’ because the simple random sample has a distribution that resembles the population.

> **Solution:** **True.** When taking a simple random sample, the shape of the distribution tends to look like the population’s distribution in shape and spread.

(7) Data on income are stored as integers, with 1 standing for the range under $50k, 2 for $50k to $80k and 3 for over $80k. This income data is quantitative.

> **Solution:** **False.** Although stored as integers, these values represent ordered categories so they are qualitative.



2. Consider the above plot about how baby boomers describe themselves. Which mistakes does it make? Circle all that apply.

   A. poor choice of color palette

   B. jiggling base line

   C. stacking

   D. jittering

   E. area perception

3. Suppose we collected purchase data consisting of **transaction id**, the purchase **amount**, and the **time of day**. If we wanted to create a visualization to explore the purchase behavior, which of the following plots would likely be helpful? Circle all that apply.

   A. a bar plot of the amount for each transaction id

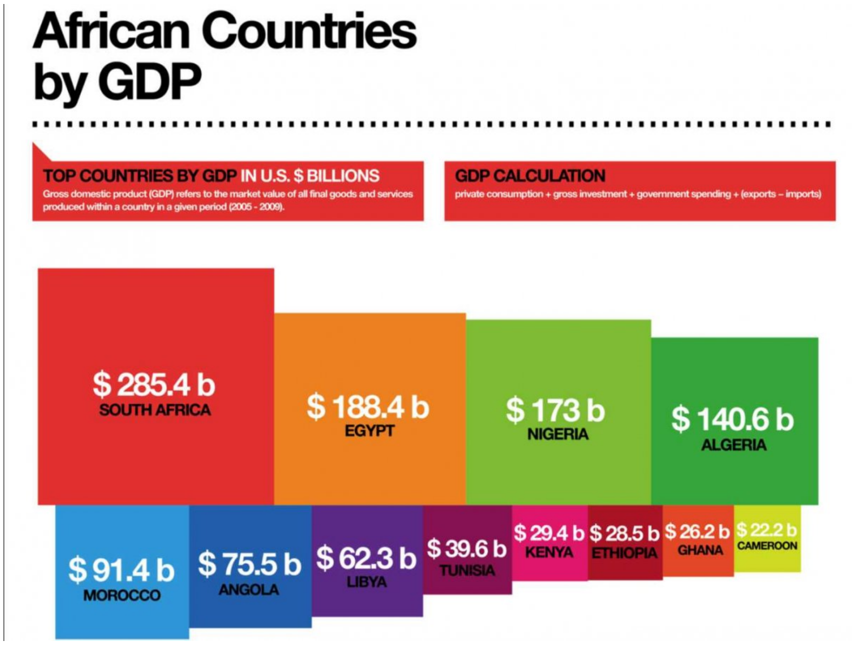
   B. density curve of transaction amounts

   C. a scatter plot of purchase amount and time of day

   D. a bar plot with the purchase for each time of day

   E. a bar plot with total purchase amount aggregated over each hour of the day.

   F. None of the above



4. Consider the figure above. Which of the following suggestions would better facilitate comparisons of the GDP for African countries. **Circle all that apply.**

   A. arrange the countries in alphabetical order to make it easier to find a country's GDP

   B. choose a sequential color palette to match size of the GDP

   C. make a box plot of GDP to show the skew and spread in GDP

   D. make a bar or dot chart of the GDP

   E. none of the above

5. Which of the following are reliable ways to assess the granularity of a table. **Circle all that apply.**

   A. Build histograms on each column.

   B. Identify a primary key.

   C. Compare the number of rows in the table with the number of distinct values in subsets of the columns.

   D. All of the above.

E. None of the above.

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

   A. $9\sigma^2$

   B. $3\sigma^2$

   Solution: This is the correct answer because variance is additive for independent random variables.

   C. σ^2

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

7. 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) What is $\mathbb{P}(X = 0)$?

   A. $1/9$

   B. $1/5$

   C. $1/4$

   D. $2/5$

   E. none of the above

   Solution: The event that $X = 0$ is the same as the event that no red marbles are drawn.

   We can use a 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 $3/15 = 1/5$.

   Alternatively, we can use conditional probability. The chance no red marbles are drawn is the same as the event that the first draw isn't red and the second draw isn't red.

$$\begin{aligned} p &= P(\text{1st draw is not red and 2nd draw is not red}) \\ &= P(\text{1st draw is not red}) \times P(\text{2nd draw is not red given 1st is not red}) \\ &= \frac{1}{2} \times \frac{2}{5} = \frac{1}{5} \end{aligned}$$

   Note that if the first draw isn't red, there are 5 marbles left, 3 of which are red.

   (2) 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. $\frac{8}{15}$

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 $2/15$.

This approach is using 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 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).

8. 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 $0 \leq p \leq 1$.

What is the expected value of X?

A. $2p(1-p)$
B. $p^2(1-p)^2$
C. 0
D. $1-2p$
E. 1

Solution: The expected value of X is

$$\begin{aligned} E(X) &= \sum_{i=1}^{m} v_i P(X = v_i) \\ &= -1P(X = -1) + 0P(X = 0) + 1P(X = 1) \\ &= -p^2 + (1-p)^2 \\ &= 1 - 2p \end{aligned}$$

9. Use the following hypothesis:

> *Berkeley students who have taken Data8 are more likely to be hired as data scientists than those who have not taken Data8.*

to answer each of the following questions. For each of the following questions **circle all of the appropriate answers:**

(1) Which of the following is the population:

A. All students in the US
B. Berkeley students
C. Students who have taken Data8
D. Berkeley students with job offers.
E. none of the above

(2) A dataset was constructed by inviting Data8 students to complete a voluntary survey. Such a dataset would most likely be described as a:

A. Sample
B. Census

(3) Which of the following are reasons the voluntary survey of Data8 students would be insufficient to make a conclusion about the hypothesis?

A. The sample size is guaranteed to be too small.
B. The survey may not be representative of Data8 students overall.
C. The survey would tell us nothing about non-Berkeley students.
D. The survey would tell us nothing about students who have not taken Data8.

E. The survey would tell us nothing about students who were not hired as data scientists.
F. None of the above.

(4) A second analysis was conducted by asking Berkeley graduates employed as data scientists. Together with the survey of Data8 students, would this be sufficient to make a conclusion about the hypothesis?

A. Yes
B. No

Solution: This problem is slightly tricky. The survey of Data 8 students would not give us any data about students that did not take Data 8. While the survey of data scientists would not provide information about students who did not become data scientists. In particular neither of these samples would contain the Berkeley students who did not take Data8 and did not get a job as a data scientist.

10. A town has 200 families, where 20% have 0 children, 30% have 1 child, and 50% have 2 children. The names of all the children are written on tickets and placed in a glass bowl. The tickets are well mixed. One ticket is drawn. What is the chance the child is from a 2-child family? Assume the children's names are unique.

    A. $1/3$

    B. $1/2$

    C. $5/8$

    D. $10/13$

    E. none of the above

Solution: We can compute the solution by looking at the fraction of tickets in the barrel that come from 2 children families. It is important to note the following two conditions

- There will be no tickets corresponding to families with no children
- There will be two tickets for each family with two children

$$\frac{200 \cdot \frac{5}{10} \cdot 2}{200(\frac{5}{10} \cdot 2 + \frac{3}{10})} = \frac{200}{260} = \frac{10}{13}$$

OR

$$\frac{\frac{5}{5+3} \cdot 2}{\frac{5}{5+3} \cdot 2 + \frac{3}{5+3}} = \frac{5 \cdot 2}{5 \cdot 2 + 3} = \frac{10}{13}$$

11. Select **all** the strings that **fully match** the regular expression: **`toy+(boat)*`**

    A. **`toy`**

    B. `toy(boat)`

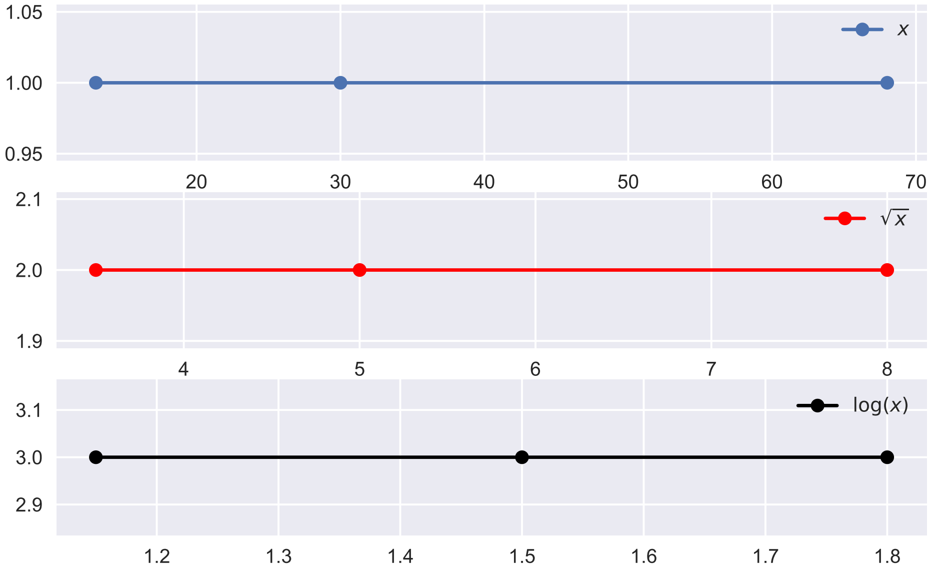
    C. **`toyboat`**

    D. **`toyyyyboatboat`**

    E. None of the above.

12. Consider the following statistics for x, which is infant mortality rate for 200 countries. According to these, which transformation would symmetrize the distribution?

Transformation	lower quartile	median	upper quartile
x	13	30	68
$\sqrt{x}$	3.5	5	8
$log(x)$	1.15	1.5	1.8



A. no transformation

B. square root

C. log

D. not possible to tell with this information

Solution: We would take a log transformation because the ratio

$$\texttt{(upperQ} - \texttt{median)/(median} - \texttt{lowerQ)} \tag{1}$$

for these 3 cases is $38/27 = 1.4$ for the untransformed data, $3/1.5 = 2$ for the square root transformation, and $0.3/0.35 = 0.86$ for the log transformation. The log transformation gives us a value closest to 1 and so is most symmetric of the possibilities.

Also, we can see from the statistics for the original data that the distribution appears skew right and the range between smallest and largest values is more than 5 so a log transformation should help make the distribution symmetric.

13. For the following population, $\{2,2,2,2,4,4,6,6,6,6,6,6,8,8,8\}$ we take a SRS and get $\{2,2,6,6,8\}$. Which of the following could not possibly be a bootstrap sample?

   A. $\{2,2,2,6,8\}$

   B. $\{2,2,6,8\}$

   C. $\{2,2,6,6,8\}$

   D. $\{2,2,4,6,8\}$

   E. All of the above are possible bootstrap samples.

> **Solution:** The sample is used as a bootstrap population, and we take a sample with replacement of 5 from the bootstrap population.
>
> Since we sample with replacement from the bootstrap population, it is possible to get three 2s in our bootstrap sample, even though the original sample only has two 2s.
>
> The bootstrap sample is the same size as the sample so it must be a collection of 5 values.
>
> Since the sample does not contain any 4s, the bootstrap sample could not have any 4s either.

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

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

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

> **Solution:** Taking the derivative of the loss function with respect to the parameter λ we get:
>
> $$\begin{aligned}\frac{\partial}{\partial\lambda}L(\lambda,\mathcal{D}) &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\log\left(\lambda e^{-\lambda x}\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\left(\log\left(\lambda\right)+\log\left(e^{-\lambda x}\right)\right) && (2)\\ &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\left(\log\left(\lambda\right)-\lambda x\right) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\lambda}-x\right) && (3)\\ &= -\frac{1}{\lambda}+\frac{1}{n}\sum_{i=1}^{n}x_i && (4)\end{aligned}$$
>
> To compute the loss minimizing parameter $\hat{\lambda}$ we set the above derivative equal to zero and

solve.

$$0 = -\frac{1}{\lambda} + \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i} \tag{7}$$

Thus the loss minimizing parameter estimate is:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1} = \frac{1}{\textbf{Mean}(x)} \tag{8}$$

15. For the following parts, please write the corresponding Python code or regular expression for the task.

(1) Write a regular expression that matches a string that contains only lowercase letters and numbers (including empty string).

Solution:

```
regx = '^[a-z0-9]*$'
```

(2) Given **`text1 = "21 Hearst Street"`**, use methods in `RE` module to abbreviate **`"Street"`** as **`"St."`**. The result should look like **`"21 Hearst St."`**.

Solution:

```
re.sub('Street',  'St.', text1)
```

(3) Given **`text2 = "October 10, November 11, December 12, January`**

1", use methods in `RE` module to extract all the numbers in the string. The result should look like **["10", "11", "12", "1"]**.

Solution:

```
re.findall(r’\d+’, text2)
```

16. For the following parts, select **all** the strings that **fully match** the regular expression:

    (1) **ab.*A**

    A. abAbA

    B. abA

    C. ab.A

    D. ab.

    E. None of the above strings match.

    (2) **ab.*?A**

    A. abAbA

    B. abA

    C. ab.A

    D. ab.

    E. None of the above strings match.

17. 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**
id				
123	4	brown	shaggy	odie
456	3	grey	short	gabe
821	6	golden	curly	samosa
198	4	grey	shaggy	gabe
3	2	black	curly	bob barker
42	5	brown	shaggy	odie

Solution: In case you want to try some of these functions in python here is the code to generate this dataframe.

```
dogs = pd.DataFrame([
    {"id": 123, "age": 4, "color": "brown", "fur": "shaggy",
        "name": "odie"},
    {"id": 456, "age": 3, "color": "grey", "fur": "short",
        "name": "gabe"},
    {"id": 821, "age": 6, "color": "golden", "fur": "curly",
        "name": "samosa"},
    {"id": 198, "age": 4, "color": "grey", "fur": "shaggy",
        "name": "gabe"},
    {"id": 3,   "age": 2, "color": "black", "fur": "curly",
        "name": "bob barker"},
    {"id": 42,  "age": 5, "color": "brown", "fur": "shaggy",
        "name": "odie"}
]).set_index(’id’)
dogs
dogs
```

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. `dogs["name"].unique().size`

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

C. `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?

A. `dogs[’age’].max()`

B. `dogs.loc[dogs[’age’].max()][’name’]`

C. `dogs.loc[dogs[’age’].argmax()][’name’]`

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

Solution: The first solution returns the age of the oldest dog. The second solution makes little sense as it uses the age of the oldest dog to lookup the row by the dog id. The fourth 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?

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.

Solution: This is a tricky question. The initial `groupby("color").count()` groups rows by color and counts the number of rows in each color. The resulting value for each column are then just the counts in each row. Therefore it doens't matter which column we use to sort.

(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?)

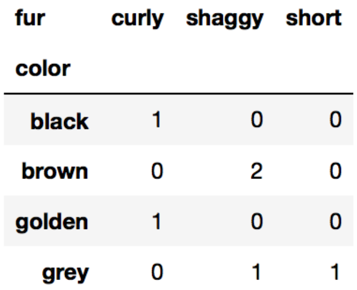
A. `(dogs['fur'].value_counts() / dogs.size)`

B. `(dogs['fur'].value_counts() / dogs.size).max()`

C. `(dogs['fur'].value_counts() / dogs.size).argmax()`

D. None of the above.

(5) Construct a DataFrame containing the number of dogs with a given color and fur type:



fur	curly	shaggy	short
color			
black	1	0	0
brown	0	2	0
golden	1	0	0
grey	0	1	1

Write the solution on the following line. You should require a single function call using a function provided on the cheat sheet.

Solution:

```
pd.pivot_table(dogs,
      index      = "color",
      columns    = "fur",
      values     = "name",
      aggfunc    = "count",
      fill_value = 0.0)
```

End of Exam