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.

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

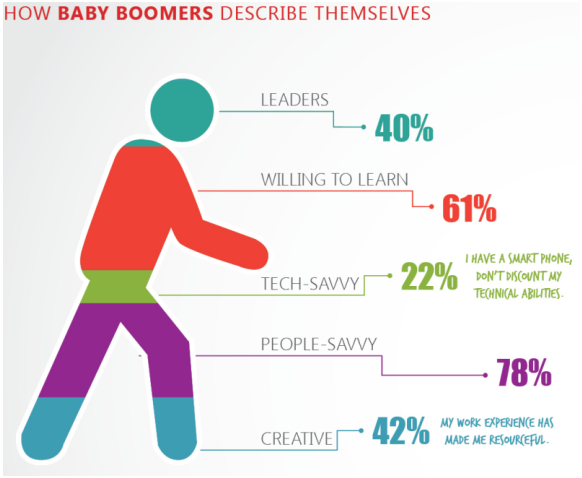
   (3) Most data scientists spend the majority of their time developing new models.

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

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

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

   (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.



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

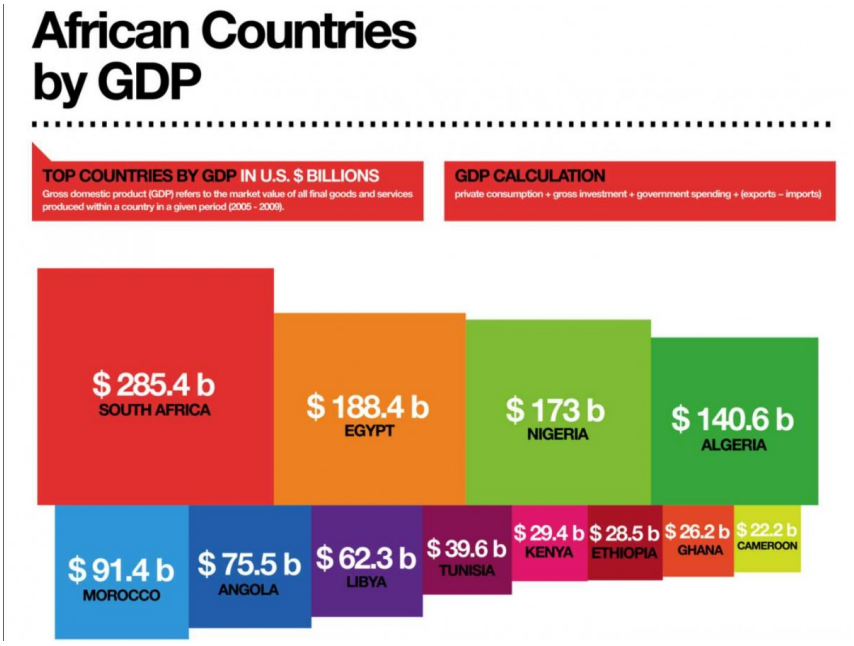
   B. density curve of transaction amounts

   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 $\mathrm{Var}(X) = \sigma^2$, then $\mathrm{Var}(X + Y + Z)$ is:

   A. $9\sigma^2$

   B. $3\sigma^2$

   C. σ^2

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

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

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

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

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

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

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

    A. `toy`

    B. `toy(boat)`

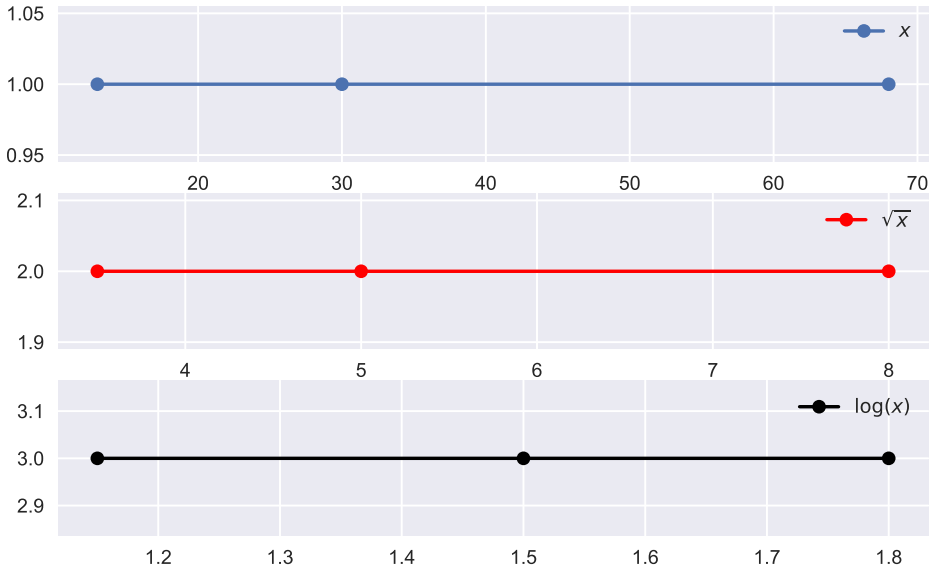
    C. `toyboat`

    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

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.

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.**

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).

---

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

---

(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"]`**.

---

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.

id	age	color	fur	name
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

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)`

(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"})`

(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['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 color	curly	shaggy	short
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.

End of Exam